CUK-H05967-14-KP2256

THE JOURNAL OF PHILOSOPHY

02/4/03

VOLUME C, NUMBER 1
JANUARY 2003

14

page

THE DEWEY LECTURES 2002:
Truth and the Past Michael Dummett

5 Lecture I: *The Concept of Truth*
26 Lecture II: *Statements about the Past*
38 Lecture III: *The Metaphysics of Time*

54 MEMO TO AUTHORS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX. (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

100
J826

KP 2256

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER. Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923. Since its founding, it has been published from the Department of Philosophy at Columbia University. It is typeset and printed by Capital City Press.

02\19.\03

THE JOURNAL OF PHILOSOPHY

VOLUME C, NUMBER 1
JANUARY–DECEMBER 2003

Edited by
Bernard Berofsky, Akeel Bilgrami
Arthur C. Danto, Kent Greenawalt
Patricia Kitcher, Philip Kitcher
Isaac Levi, Mary Mothersill, Philip Pettit
Carol Rovane, Achille C. Varzi

Consulting Editors
David Albert, John Collins
James T. Higginbotham, Charles D. Parsons
Wilfried Sieg

Managing Editor
John Smylie

Published by The Journal of Philosophy, Inc.

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all photocopying. Inquiries about photocopying may be sent to the Managing Editor.

© Copyright 2003 by the Journal of Philosophy, Inc

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and other mailing offices. Please send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. Fax (212) 932-3721

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 1, JANUARY 2003

TRUTH AND THE PAST

LECTURE 1: THE CONCEPT OF TRUTH

The distinction between linguistic utterances and what they express is borne in on us by many common experiences—for instance, that of finding out how to say something in another language, and that of rephrasing something we have said to make it clear to our hearer. What a sentential utterance expresses is a proposition. A question frequently discussed is whether 'true' is to be taken as predicated primarily of propositions or of sentences considered as uttered on particular occasions.

In my view, this is not a deep question. The reason for taking 'true' to be predicated of linguistic items is that we need to appeal to the notion of truth in order to recognize what proposition a given declarative sentence expresses. This is shown by the evident cogency of Gottlob Frege's necessary condition for two sentences to express the same proposition, namely, that it should be impossible for anyone who understands both to conceive of one's being true and the other not. If we regard truth as attaching to propositions rather than to sentences expressing them, we are taking the concept of a proposition as prior to that of truth, and debar ourselves from using the concept of truth to explain when two sentences express the same proposition; it then becomes doubtful whether we can explain this at all, and hence what a proposition is.

Frege repeatedly declared that it is of propositions (thoughts, in his terminology) that truth and falsity are predicated; but his practice, in the *Grundgesetze der Arithmetik*,[1] proceeds in the opposite direction.

---

[1] Jena: Pohle, 1893–1903. Translated in part as *The Basic Laws of Arithmetic: Exposition of the System*, Montgomery Furth, ed. and trans. (Berkeley: California UP, 1964).

0022-362X/03/0001/5–53

He lays down stipulations intended to determine the reference (*Bedeutung*) of every expression, simple or complex, of his formal system. Frege's notion of *Bedeutung* is not to be interpreted as "denotation," but, rather, as "semantic value"—the contribution an expression makes to determining the truth or falsity of a sentence in which it occurs, the *Bedeutung* of a sentence being its truth value. Frege's stipulations thus serve to determine the truth value each formal sentence has: and they do so independently of the notion of sense, and in advance of its being introduced at all. The stipulations lay down, inductively, what the semantic value of each expression is to be: the sense of an expression consists in the way in which these stipulations combine, in accordance with how that expression is made up out of its parts, to determine it as having a certain semantic value. The thought expressed by a sentence therefore consists in the condition, as given by those stipulations, for it to have the value *true*. To arrive at this account, it was necessary to take the notion of a sentence's having the value *true* as prior to that of the thought it expresses.

Why, then, should Frege and others after him have insisted that truth attaches to propositions and not to a piece of language? This is prompted by our practice in judging of truth and falsity. Notoriously, the occurrence of indexical expressions prevents us from taking truth to be an attribute of type sentences of natural language; but indexicality is far from being the only bar to this. Even given the denotations of the indexical and demonstrative expressions they contain, we have no formal means of circumscribing those utterances which we treat as apt for evaluation as true or false, either by their form or by their context. We must exclude sentences devoid of truth value because of not expressing any intelligible proposition at all, for instance, the sentence 'Hydrogen is loud'; but even this is not easy to do. In general, we must pick out only forms of words for which it is for us determinate in which circumstances what they say may be recognized as true; those, that is, which we understand as expressing specific propositions. We cannot exclude all sentences containing ambiguous or vague expressions, or those used metaphorically or loosely. In certain contexts, no competent speaker would understand an utterance in more than one of the senses it could, in principle, bear. Often, the vagueness of an expression is no obstacle to our pronouncing a statement containing it to be certainly true or certainly false. Some metaphors are dead or familiar enough to be recorded in dictionaries. Predicates may have a quite determinate theoretical application, yet often be legitimately used in a much looser sense: an excellent example of

Peter Unger, repeated by the late David Lewis,[2] is the adjective 'flat'. It depends on the context how a statement involving any of these features is to be understood; the context cannot be readily specified, however, but depends on what the speaker is likely to have thought or said.

How can we ever have arrived at the idea of propositions that are determinately either true or false? We do so because we are habituated to understanding what others say in a particular way: not by grasping all its possible meanings, according to the senses of the words in the language, but by taking it in this way or that on a particular occasion. We ignore one possible meaning of a sentence that is, in principle, ambiguous in favor of another; we treat the word 'flat' as applicable in a manner that accords with the circumstances; we correctly construe an unusual turn of phrase as readily as if it were a standard idiom. In short, we put a particular *interpretation* on it. Sometimes we need to reflect in order to hit on the meaning the speaker probably intends; more often, we adopt an interpretation without thinking or striving to attain it.

This is not the sense of "interpretation" that figures in Donald Davidson's[3] account of linguistic communication. A Davidsonian interpreter fashions an entire theory of meaning for the speaker's language: an interpretation in the present sense is the selection of one meaning out of several that the language will allow the sentence to bear. The senses of the words of a natural language are extremely flexible. When we put them together in a sentence, these senses are greatly stiffened by the presence of the other words; they are stiffened still more by the whole context in which the sentence is uttered. Successful communication depends on its being our habitual practice so to construe an assertoric utterance as to render it as determinate as possible which circumstances will warrant its being judged true and which its being judged false: that is, as expressing a definite proposition. We normally do this automatically, without having to review the different interpretations that the language would admit as possible. It is through our familiarity with this process that we come to grasp the concept of a proposition.

---

[2] See Unger, *Ignorance: A Case for Skepticism* (New York: Oxford, 1975), Chapter 2; Lewis, "Scorekeeping in a Language Game" (1979), reprinted in *Philosophical Papers*, Volume I (New York: Oxford, 1983), pp 233–49, and "Elusive Knowledge" (1996), reprinted in *Papers in Metaphysics and Epistemology* (New York: Cambridge, 1999), pp. 418–45.

[3] For example, "Radical Interpretation," in *Inquiries into Truth and Interpretation* (New York: Oxford, 1984), pp. 125–39.

We should thus see truth as attaching to a token sentence, but one considered under a particular interpretation. The language determines what interpretations are in principle possible; the circumstances of utterance determine which interpretation is obviously intended. The sentence so interpreted will probably not have a truth value in every imaginable situation; with luck, it will have one in every situation that actually arises. It will therefore not express a thought according to Frege's strict criteria; but we may take it as expressing a proposition that can legitimately be said to be true or false.

By far the most popular type among theories of meaning is the truth-conditional variety, according to which the core meaning of a statement—its *sense*, in Frege's terminology—is given by, or consists in, the condition for it to be true. There are two parts to knowing what a word or sentence means: we have to grasp the concept or proposition it expresses; and we have to know that that *is* the concept or proposition it expresses. Any account purporting to provide a full explanation of the meaning of the word or sentence must cover both these aspects. There are two opposite errors that reduce an explanation of meaning to only one of these two components. The proponents of what are known as *modest* theories of meaning maintain that it is no business of a theory of meaning to explain the concepts expressed by the words of a language, but only to specify which concepts they express. Such a theory is not merely modest, but threadbare: by failing to explain what is needed for someone to grasp the concept expressed by a word, it omits an essential ingredient of the understanding of the word.

The opposite error is most likely to be made in application to whole sentences; and it is usually advanced by those who display an insouciant attitude to quotation marks. The notion of truth, they argue, is needed only to state the *general* principle that to know the content of a proposition is to know the condition for it to be true: in any particular application the notion of truth is redundant. This thesis can hold, at best, when we are seeking to explain what it is to know the content of a proposition, not when we want a complete explanation of what it is to understand a linguistic utterance. It is true that, in conversing or reading in any language that we know well, we do not carry out two successive activities: first listening to a sequence of sounds or looking at a string of marks on paper, and then interpreting them. Rather, as Edmund Husserl[4] insisted, we hear or read the words *as* saying whatever it is that they say; only by a heroic effort can we hear them as mere sounds or see them as mere shapes. We perceive the meaning *in* the words. In the same way, when we are traveling in a country whose language we do *not* know, but which uses the Roman script with which we are familiar, we cannot help reading words and names that we see written up as being pronounced in a certain way, although we know that the natives' pronunciation may be quite different. But the psychological fusion of spelling and sound does not obliterate either the distinction between them or our ability to separate them in thought; likewise, the psychological fusion of verbal form and sense does not obliterate either the distinction between *them* or our ability to separate them in thought. The distinction between the linguistic vehicle and the proposition it is used to express is not to be repudiated on phenomenological grounds such as these.

---

[4] *Logical Investigations*, Dermot Moran, ed., J. N. Findlay, trans. (New York: Routledge, 2001), Volume I, Investigation I, §10, pp. 193–94.

The argument that the notion of truth is dispensable in particular cases is plausible only when an inquiry after the meaning of a sentence of some language is to be answered in that language. If someone, reading a book, asks me, "What does 'fratricide' mean?" I might reply, "To commit fratricide is to kill one's brother": my reply does not refer to the word 'fratricide', but only uses it. You could not, however, explain in English what the Chinese or Arabic term for 'fratricide' meant without mentioning that term. Likewise, to know the condition that must hold good for swans to sing before they die is only one of the things I must know in order to know what a sentence translatable into English as 'Swans sing before they die' means. If the explanation is to involve the notion of truth, truth must be treated as an attribute of a sentence under a particular interpretation, rather than just of the proposition it expresses when so interpreted. For a truth-conditional theory of *meaning*, the notion of truth as an attribute of sentences is indispensable.

Ought meaning be explained truth conditionally? Ludwig Wittgenstein thought, or at least frequently wrote as if he thought, that the meaning of a word or type sentence is constituted by its use, that is, the manner in which it is (correctly) employed in linguistic interchange (spoken or written). On the face of it, this seems undeniable. How else do we tell whether or not someone has mastered a language than by attending to the use he makes of its words and sentences—observing what he says when using that language, how he responds when someone speaks in that language to him, and other features of his associated behavior? If we want to engage in what Davidson calls radical interpretation—trying to understand the speakers of a language unassisted by any bilingual interpreter, phrase book, or dictionary—what else is there to go on but the use the speakers make of its words and sentences?

What constitutes "use"? Forms of sentence commonly used to express something different from what their composition would suggest—for instance, tautologies such as 'Boys will be boys' and 'What is done is done'—illustrate only a peripheral feature of use. In Wittgenstein's writings, there are three fundamental features of use. First is what is accepted as justifying an utterance. If the utterance is of an assertoric character, this becomes what vindicates it or establishes it as correct. This comes under the head of the *grounds* for asserting a statement: these include what is acknowledged as entitling a speaker to make the assertion, what requires him to withdraw it, and in what circumstances, if any, it is conclusively shown to be correct.

The second feature consists of what a speaker commits himself to by saying something, and what is counted as a proper response to another's utterance. When the utterance is assertoric, this comes under the head of the *consequences* of making the assertion: what actual or potential difference it makes to what is done or said by the speaker or his hearers. It includes what constitutes acting on an assertion when it is accepted, as well as inferences drawn from the statement so asserted.

The third feature is a more reflective one, namely, the point of having a given form of words in the language. To ask after its point is to ask what we are able to do with that form of words which we could not do without it. It may seem worth while to inquire, for instance, what is the point of having such a form as counterfactual conditionals, and right to criticize some account of their meanings on the ground that it deprives them of any point. But, although a question of the form, 'What is the point of having such-and-such a form of words in the language?', is often philosophically useful to ask, it is arguable that this is not a genuinely independent feature of use: a fully adequate account of the use of a given form of sentence will always *show* what is the point of having it in the language, whenever there really is a point in this.

None of these features of a language involves the conception of a statement's *being* true. The first two of them, when applied to the making of statements, may be said to be concerned with notions that involve truth: that of establishing as true, and that of acting on the truth of something. So could we explain meaning in terms of one or other of these notions, or perhaps of both of them together? By far the most systematic attempt to do so, although only for a restricted segment of language, is the intuitionistic theory of meaning for mathematical statements. In this, the meaning of a mathematical statement is not explained in terms of the condition for it to be true, independently of our knowledge. It is explained, rather, in terms of what is required of a proof of that statement. A proof is what is needed in order to warrant the assertion of a mathematical statement: the intuitionist account of meaning thus explains the statement in terms of the first of the three foregoing features of its use. The consequences of accepting a mathematical statement as true are that we can use it to make inferences that lead to proofs of further statements. If we know the meanings of all other mathematical statements, as given in these terms, we shall also know what is required for drawing a conclusion from any given statement; so we do not need a separate specification of the second feature of its use.

It is plausible that this latter point can be generalized. Wittgenstein, engaged in informal discussions of one or another form of words, sometimes concentrates on one of the two fundamental features of use, sometimes on the other, sometimes on both of them simultaneously. But when we are aiming at sketching the form that a systematic theory of meaning for a whole language ought to take, and are aiming to formulate the theory directly in terms of one or more features of the use of statements of that language, it is plausible that we may frame it in terms of just one of the two fundamental features of use, just as the intuitionist theory of meaning for mathematical statements is framed in terms of what is needed for the proof of a given statement, and not also in terms of what could be proved from it. This is because it is a requirement upon a language that functions as a language ought that there be harmony between the two features of use: that is, that we draw from a statement only consequences that match the grounds for asserting or accepting it, and that we accept it on any grounds that justify the consequences we draw from it. If such harmony prevails, then one of the two fundamental features of use will be derivable from the other; and the meanings of individual words and sentences of the language can be given in terms of either such feature. A theory of meaning given in terms of the grounds for asserting a statement I shall call a *justificationist* theory; one given in terms of the consequences of accepting a statement I shall call a *pragmatist* theory. The intuitionist theory is a justificationist theory of meaning for mathematical statements.

There is indeed no guarantee that the linguistic practices accepted by the speakers of any actual language will conform to the requirements of harmony. But, if they do not, those practices are in need of reform. A theory of meaning for a language in need of reform will probably be imperfectly coherent, and will in any case be of minor value. If the best fully coherent theory of meaning for a language fails to fit completely with the conventional practices of its speakers, the language is in need of reform, and the theory will show in which respects it needs to be reformed.

A proponent of a truth-conditional theory of meaning is not aiming to give a stipulative definition of "meaning": he is not just laying down that, for him, the meaning of a statement is constituted by the condition for it to be true. If he were, his thesis would be trivial. Rather, he is aiming to *explain* what meaning consists in. Now, what anyone needs to know in order to be able to engage in the communal practice of speaking and writing a particular language is the use that is made, by those who know it, of its expressions and sentences; and, as already remarked, the criterion for his knowing the language is that he manifest such knowledge. The truth-conditional theorist therefore has two challenges to meet. He has, first, to show how the fundamental features of the use of any statement of the language can be derived from the conditions for its truth: how we can from these conditions determine what are the grounds for asserting the statement, and how accepting it will affect the actions of a speaker or hearer. In showing this, the theorist will be spelling out that connection between the use of a language and the truth conditions of its statements—the connection between truth and meaning—which Davidson, in his original formulations, left as tacitly understood. Until he has done this, the truth-conditional theorist will not yet have vindicated his claim that his theory of truth conditions constitutes a theory *of meaning*. To vindicate it, he admittedly need not derive all aspects of the use of a statement: he may put some down to particular linguistic conventions. For instance, he may reasonably hold that there are conventions governing entitlement to assert statements of different forms, conventions that do not affect the meanings of those statements. We do not regard it as proper to assert a mathematical statement unless it has been proved; otherwise, it can only be conjectured. We can easily imagine, however, people who regard it as legitimate to assert such a statement when there is a probable ground, falling short of proof, for conjecturing its truth, even though they have the same standards of proof as we do and are as interested in discovering proofs: they would mean the same by their mathematical statements as we mean by them. But the theorist must be able to determine from the condition for the truth of any statement what—if anything—counts as *conclusively* establishing it.

By meeting this first challenge, the truth-conditional theorist will have shown that his theory is truly a theory of meaning; but he will not have shown that a theory of meaning must necessarily assume a truth-conditional form. After all, if the criterion for knowing a language is that a subject manifests a knowledge of the use of its expressions and sentences, will not the most direct theory of meaning for that language be a straightforward (though, of course, very complex) description of how the language functions, that is, an account of the use of its words and sentences? Why do we need a digression via the truth conditions of its statements, even if this eventually yields the account of use that we are after? To demonstrate the necessity of a truth-conditional theory of meaning, a proponent of such a theory must argue that use cannot be described without appeal to the conditions for the truth of statements; in so far as he is aiming to explain understanding as well as meaning, as he ought to be aiming to do, he must claim that anyone who has mastered the use of expressions of a language must have acquired an implicit grasp of the concept of truth and an implicit conception of the conditions for the truth of its statements. It will then follow that a theory of meaning in terms of use must be a truth-conditional theory.

The tendency is widespread among philosophers to regard as inviolable the principle of what, following W. V. Quine,[5] may be labeled *semantic shift*. By an instance of this principle I mean a biconditional connecting a declarative sentence with another ascribing truth to the statement or proposition expressed by that sentence. (An instance of semantic *ascent* is a conditional whose consequent ascribes truth to the statement or proposition expressed by the antecedent; an instance of semantic *descent* is the converse of such a conditional.) The principle may be applied to statements, considered as linguistic items, and is then conventionally illustrated by the biconditional:

> The statement 'Snow is white' is true if and only if snow is white.

Some prefer to apply it to propositions, as illustrated by:

> The proposition that snow is white is true if and only if snow is white.

This may be more simply framed as:

> It is true that snow is white if and only if snow is white.

For this third biconditional to be a formulation of semantic shift, the 'that'-clause following 'true' must be understood as denoting a proposition; if the expression 'it is true that' is taken as a sentential operator carrying any sentence into one with the same truth value, it will involve no semantic shift whatever.

---

[5] *Philosophy of Logic* (Englewood Cliffs, NJ: Prentice-Hall, 1970), see p. 12.

The principles of semantic shift are very far from being inviolable. The simplest demonstration of this is to apply them to sentences that generate semantic paradoxes of the liar type. The resulting antinomies cannot be evaded by declaring that the paradoxical sentences express no propositions: if we have the liar say, 'I am now expressing no true proposition', the claim that he is not thereby expressing a proposition appears wholly arbitrary. It is much better to recognize that, in such a case, the principle of semantic shift fails. In saying, 'I am now expressing no true proposition', the liar was expressing a proposition, but not a true one; but we cannot, by invoking the principle of semantic ascent, infer from that fact that he *was* expressing a true proposition.

We saw that the proponent of a truth-conditional theory of meaning must argue that use cannot be described without appeal to the conditions for the truth of statements, and that anyone who has mastered the use of expressions of a language must have acquired an implicit grasp of the concept of truth. Such an argument would in considerable part be correct. An adequate theory of meaning must yield an account of what difference the assertion of a statement makes, actually or potentially, to what subsequently happens. This includes the difference that acceptance of the statement by a hearer makes to what he subsequently thinks, says, and does. But what, in general, does accepting some statement as true involve? We do not merely react piecemeal to what other people say to us: we use the information we acquire, by our own observation and inferences and by what we are told, in speech and writing, by others, to build up an integrated picture of the world. To take a statement as contributing to this picture of the world, that is, of reality independent of our will, is to take it as true. So our practice of using language does impart to us an implicit grasp of the concept of truth.

The intuitionist account of the meaning of mathematical statements does not employ the notion of a statement's being true, but only that of something's being a proof of the statement. It is therefore natural to conclude that a justificationist theory of meaning can dispense with the notion of truth—that is, of a statement's *being* true, as opposed to its being recognized or accepted as true—altogether; presumably the same will apply to a pragmatist theory of meaning. This is a mistake, however. That part of the intuitionist theory which specifies the meanings of particular expressions uses the notion of proof, but not the notion of truth; but, as an overall theory of meaning for mathematical statements, it needs a notion of truth. This is for two reasons, which apply to any theory of meaning whatever.

If Frege's account of sense is right, as I believe, then the theory of sense rests upon a semantic theory as base. Frege explained the sense of an expression as the way in which its semantic value is given to us. His semantic theory was a classical two-valued one: so, for him, the semantic value of an expression was its contribution to determining the truth value of a sentence in which it occurred. This agrees with the way in which a semantic theory is usually characterized—it explains how the truth value of a sentence is determined in accordance with its composition. But many theories of meaning do not take a statement's simply being true or not being true as their central notion. The intuitionistic theory takes as its central notion that of a construction's being a proof of a statement—a choice that shrinks the gap between sense and semantic value to the minimum; a theory that rests on the possible-worlds semantics for statements involving modality takes as its central notion that of a statement's being true at a possible world. We must therefore generalize our characterization of a semantic theory. In any theory of meaning, a statement will have what may be called a *fundamental semantic property*, definable in terms of the central notion of the theory. In a possible-worlds theory, this will be the property of being true at the possible worlds in some set and not at others; in the intuitionistic theory, it is that of being proved by certain constructions and not by others. A semantic theory must then be characterized as one that explains how the fundamental semantic property of any statement is determined in accordance with its composition. We may still hold that, within any theory of meaning, its component theory of sense will rest on the base of a semantic theory, understood in this generalized manner.

Now, a comprehensive theory of meaning for any language must aim at giving a complete account of how that language functions. The philosophical purpose of considering the form that a theory of meaning for a particular language should take does not lie in any special interest we might have in that individual language, but is to illuminate the general nature of linguistic meaning: what it is for linguistic utterances to have the significance that they have, and what, therefore, it is to have a language at all and to engage in the practice of speaking it. Hence a theory of meaning for a particular language should be conceived by a philosopher as describing the practice of linguistic interchange by speakers of the language without taking it as already understood what it is to have a language at all: that is what, by imagining such a theory, we are trying to make explicit. A salient part of using a language is to give arguments in support of some conclusion, either to induce the hearer to accept that conclusion or to justify a prior assertion of it when challenged. Hence such a theory of meaning ought, as one of its tasks, to give an account of the procedure of deductive inference, and supply a criterion for the validity of such an inference. A semantic theory is used by logicians for just this purpose; it explains how the fundamental semantic property of any statement is determined in accordance with its composition in order thereby to characterize valid deductive inferences. It is here presupposed that the validity of an inference will always turn solely on the semantic values of the expressions involved; it is assumed that any inference whose validity depends upon their specific senses can be handled by replacing such expressions with equivalents as laid down by their definitions.

A deductive argument is valid if it is guaranteed to transmit some desirable property from premises to conclusion: this property is normally taken to be truth. A theory of meaning, and indeed a semantic theory, thus needs a notion of truth, as that which is guaranteed to be transmitted from premises to conclusion of a deductively valid argument, in addition to the central semantic notion, if that is not already the notion of truth. But the point is more general than that. What is the meaning of a declarative sentence? One answer might be: it is the principle that governs what it serves to convey to a hearer when that sentence is used on its own on any occasion to make an assertion, that is, how the hearer takes things to be if he accepts the assertion as correct. This is indeed an important feature of the meaning of the sentence, and it is how the question what the sentence means is often answered. But it is only one feature of the sentence's meaning: we may call it the *assertoric content* of the sentence. It plainly does not constitute the whole of what the sentence means. We need to know, in addition, what contribution that sentence makes to the assertoric content of a more complex sentence of which it is a subsentence, and this is not in general determined by its own assertoric content. We may call this second feature the *ingredient sense* of the sentence. Two sentences may have the same assertoric content, but different ingredient senses. Examples may most easily be given by exploiting the phenomenon called "rigidity" by Saul Kripke.[6] The two sentences

---

[6] *Naming and Necessity* (Cambridge: Harvard, 1980), see p. 48.

It is raining here.

and

It is raining where I am.

have the same assertoric content: if you believe a friend who, speaking to you on the telephone, utters either sentence, you learn exactly the same as if he had uttered the other sentence of the pair. But the sentences do not have the same ingredient senses, as is shown by the quite different meanings (assertoric contents) of the two sentences that result from inserting the quantifier 'always':

It is always raining here.

and

It is always raining where I am.

The divergence occurs because the adverb 'here' is temporally rigid, while the adverbial phrase 'where I am' is temporally flexible. There are examples of a different kind. The sentences

I shall give you a D.

and

I intend to give you a D.

have the same assertoric content; but their ingredient senses differ, since the conditionals

If I give you a D, you will forfeit your grant.

and

If I intend to give you a D, you will forfeit your grant.

have different assertoric contents. (In English, the antecedent of the first conditional is grammatically in the present tense, but its sense is future: in Italian, it would have the same future-tense form as the simple sentence.) Sentences carrying presuppositions provide further examples. The sentences

Clare's husband is Egyptian.

and

Clare is married to an Egyptian.

have the same assertoric content, since a hearer who did not know that Clare was married would learn this as well from hearing the first sentence as from hearing the second. But they do not have the same ingredient sense, since their negations are not equivalent (negation does not cancel presupposition). One way to understand the traditional semantics for many-valued logics, with its distinction between designated and undesignated values, is to take the assertoric content of a sentence to be given by the condition for it to have a designated truth value, while the distinctions between different undesignated values, and those (if any) between different designated ones, serve to explain the ingredient senses of sentences.

The notion of assertoric content provides a second reason why a theory of meaning needs a concept of truth. Central to any description of the practice of using a language is an account of its use to make assertions and of the responses they evoke in their hearers. The description must cover other uses of language, but assertion is primary: questions call for assertions to be made, and it must be possible to say whether a request has been granted or declined, a demand complied with or flouted. It is from the practice of making assertions that the notion of truth first arises: it is of the essence of that practice that assertions can be judged correct or incorrect. We may accordingly take a statement to be true if it would be correct to assert it; more exactly, if an assertion of it would be justified, whether or not a particular speaker would have been justified in making it. This obviously depends on the assertoric content of the sentence used to make the statement, since it relates to the import of that sentence uttered on its own. The property of statements thus characterized coincides with the property guaranteed to be transmitted from premises to conclusion of a valid deductive inference: for what we need to know, of such an inference, is that, if we are entitled to assert the premises, we shall be entitled to assert the conclusion.

The defenders of truth-conditional theories of meaning are thus right to argue that the concept of truth is indispensable. The concept of truth is the pivot about which a theory of meaning is brought to bear on metaphysics. Metaphysics is concerned with the general nature of reality, and, as the opening remarks of Wittgenstein's *Tractatus*[7] state, reality is constituted, not by the totality of objects that exist, but by the totality of facts that obtain. Facts are true propositions: so metaphysics concerns itself with what truths hold good in general.

---

[7] *Tractatus Logico-Philosophicus*, D. F. Pears and B. F. McGuinness, trans. (New York: Humanities, 1961).

But we can by no means conclude to the correctness of a truth-conditional theory of meaning. We cannot do so for two reasons. First, there is no argument to show that truth needs to be taken as the central notion of the semantic theory underlying a sound theory of meaning. This central notion must be selected in order to explain in terms of it the whole meanings of sentences—not just their assertoric contents, but their ingredient senses; indeed, it is especially concerned with their ingredient senses, since one of the primary functions of a semantic theory is to explain the logical constants.

Second, different theories of meaning have different conceptions of truth. It is not enough for the truth-conditional theorist to argue that we need the concept of truth: he must show that we should have the same conception of truth as he has. We need the concept of truth for the two purposes I have stated: but giving these two purposes does not yet say what it is for a statement made by means of a sentence whose meaning has been given in terms of the central notion of the theory to be true. We shall expect that any given theory will explain this—will characterize truth—by reference to its central notion. The result will not, in general, conform to the principle of bivalence upon which truth-conditional theories of meaning depend, the principle, namely, that every statement devoid of ambiguity and vagueness is determinately either true or false. For instance, the central notion of the intuitionistic theory of meaning for mathematical statements is that of proof; and within that theory, the truth of such a statement may be said to consist in the existence of a means, effective in principle, for finding a proof of it. The canons of correct reasoning in intuitionistic mathematics debar us from restricting true statements to those we have actually proved. Thus the semantic theory stipulates that a proof of a disjunctive statement '*A* or *B*' must consist of a proof either of *A* or of *B*. But we are entitled to assert such a statement even when we do not have a proof of either disjunct, but merely have an effective means in principle of finding a proof of one or other: for instance, when *A* is the statement that some large natural number is prime, and *B* the statement that it is composite. That is the notion of truth appropriate to the intuitionistic theory of meaning: it quite obviously does not respect bivalence.

We are now in a position to evaluate the principle of semantic shift. It follows from the characterization of assertoric content that, for any statement *A*, the assertoric content of *A* and of the statement '*A* is true' must be the same. It does not in the least follow that their ingredient senses must coincide. Now, the principle of semantic shift, applied to a statement *A*, is a biconditional of which *A* is one component. It therefore involves, not merely the assertoric content of *A* and of '*A* is true', but their ingredient senses. We accordingly have no general reason, independent of the particular theory of meaning that we favor, for regarding the principle as holding good. Whether it holds good in any particular case depends on our theory of meaning and the conception of truth appropriate to it, as well as on the particular statement to which they are being applied. The widespread obsession with the principle of semantic shift as known to be inviolable in advance of any further inquiry into the concept of truth is utterly misplaced.

Consider the conception of the truth of statements in the future tense according to which they become true or false only at the time to which they relate; or the more sophisticated conception under which determinism is rejected and a future-tensed statement is true at a given time only if it holds in all subsequent courses of events causally compatible with how things are and have been up to that time, and false only if it fails in all of them. These conceptions will each accord with the account that its proponent gives of the meanings of sentences in the future tense. Under the first of these conceptions, the statement 'In 2003, Bush will veto a bill outlawing the death penalty' is neither true nor false now; not being a determinist, I imagine that the same holds good under the second of the two conceptions, too. Now, an application of the principle of semantic shift yields the conditional:

> If in 2003 Bush vetoes a bill outlawing the death penalty, the statement 'In 2003, Bush will veto a bill outlawing the death penalty' is true.

By contraposition, this yields:

> If the statement 'In 2003, Bush will veto a bill outlawing the death penalty' is not true, then Bush will not in 2003 veto a bill outlawing the death penalty.

100
7826

Clearly, the proponent of neither of the two conceptions of truth will accept this, nor, accordingly, the principle of semantic shift as applied to statements in the future tense. If you believe in the principle, you will persuade them to accept it only if you can induce them to repudiate their conceptions of truth, and hence to modify their theories of meaning.

The example is a special case of the difficulty created by a semantic theory under which certain statements—for instance, ones containing proper names that may turn out to be empty—may in particular circumstances be neither true nor false. In such a theory, a statement *A* and the statement '*A* is true' form a perfect pair of statements having the same assertoric content but different ingredient senses. I do not deny the possibility of introducing a sense of 'is true' under which '*A* is true' behaves in all respects exactly like *A*. If you, believing that there was no such person as King Arthur, can answer the question, 'Did King Arthur defeat the Saxons?', only by saying, 'There was no such person as King Arthur', then, when 'is true' is used in *this* sense, and 'is false' correlatively, you can answer the question, 'Is the statement "King Arthur defeated the Saxons" true or false?', only by saying, 'There was no such person as King Arthur'. I say only that this is not the sense of 'true' needed to frame the semantic theory, nor that which answers to our intuitive ideas concerning truth. Nor, therefore, is it the sense of 'true' needed within the theory of meaning associated with that semantic theory, nor, accordingly, that which should concern philosophers.

It is because they have correctly perceived that the assertoric content of '"Snow is white" is true' or 'It is true that snow is white' and of 'Snow is white'—what is conveyed to a hearer by means of an assertion of one sentence or the other—must coincide that philosophers have been tempted to regard the principle of semantic shift as inviolable. Frege was led to say that the thought expressed by the sentence containing the word 'true' was the very same as that expressed by the sentence not containing it. If he had meant no more than that they had the same assertoric content, he would have been right. Unfortunately, he identified the sense of a sentence with the thought it expressed, and so declared the *senses* of the two sentences to be the same. He had some excuse for doing so, in that, in his semantic theory, a sentence—except when it occurs in indirect speech—contributes to the truth value of a more complex sentence of which it is part solely in virtue of being itself true or not being true, so that its ingredient sense is determined by its assertoric content; but in a more general context, independent of the semantic theory, it is a false conclusion.

Philosophers have a strong, and sound, intuition that the concepts of truth and meaning are inextricably linked. The two concepts must be explained *together*: neither can be taken as given in advance of the other, so that the other can be explained in terms of it. The concept of truth has its home within a theory of meaning, within which it is an essential theoretical notion. It cannot be explained by stipulating that the principle of semantic shift is to hold universally. That principle is to be judged as holding or not holding in the light of our conception of truth; and our conception of truth is responsible to our theory of meaning for our language.

KP 2256

The general belief that the principle of semantic shift is inviolable has had two conflicting consequences. One is the strong appeal of the minimalist account of the concept of truth. The other is the widespread acceptance of truth-conditional theories of meaning. The minimalist thesis is twofold: (1) that every instance—or almost every instance—of the principle of semantic shift, either as applied to statements or as applied to propositions, holds good; and (2) that by stipulating that all those instances hold good, we give the *whole* explanation of the meaning of the word 'true'. Both minimalist conceptions of truth and truth-conditional theories of meaning derive their attraction from a common source; but they are incompatible with one another. Minimalism conflicts with a truth-conditional theory of meaning because it assumes meaning as given antecedently to the notion of truth. This is most obvious for that version of minimalism which regards truth as primarily an attribute of propositions. Obviously, you cannot know what proposition a statement expresses unless you know what it means. What proposition is expressed by a statement depends upon that aspect of the statement's meaning that Frege called its "sense," and not on those aspects which do not affect the truth or falsity of what is said, such as that which differentiates the words 'dead' and 'deceased'. The notion of the meaning of a statement is thus prior to that of a proposition; and a minimalist of this variety is taking the notion of a proposition as prior to the notion of truth. But a truth-conditional theory of meaning explains meaning in terms of truth, and thus takes truth as prior to meaning. The two theories are thus irreconcilable.

The minimalist stipulates, by some means, instances of semantic shift, such as

> The proposition that swans sing before they die is true if and only if swans sing before they die.

The truth-conditional theorist lays down that the sense of a statement in the language for which he is giving a theory of meaning is the condition for it to express a true proposition. Suppose that that language is Arabic. How is the theory to determine which proposition is expressed by the Arabic translation of the sentence 'Swans sing before they die'? What other criterion can the theorist adopt, if he is to be faithful to his truth-conditionalist principles, than that it expresses that proposition the condition for whose truth coincides with the condition for the truth of the sentence? So he cannot avoid using 'true' as applying to sentences as well as to propositions. Hence, in order to identify the proposition expressed by the Arabic sentence in question, he will have to be able to derive from his theory of meaning that the condition for its truth is that swans sing before they die. Hence, while he might be prepared to concede that, collectively, the principle of semantic shift accounts for the notion of truth as applied to propositions, he needs in addition a notion of truth as applied to sentences (if necessary, interpreted in some way) in order to determine which proposition each sentence, thus interpreted, expresses; and this notion, being prior to that of a proposition and of the meaning of a sentence, he cannot derive from the principles of semantic shift as enunciated by a minimalist of this variety.

The minimalist cannot take the same route, however. It will do him no good to choose to stipulate the principle of semantic shift as applied to sentences or statements as well as to propositions, and then to invoke the truth-conditional criterion to determine the proposition expressed by the statement 'Swans sing before they die': that will not help him at all to determine what proposition is expressed by the translation of that statement into Arabic.

But would not that strategy work if the language in question were not Arabic but English? What is special about English is that we are taking it as the metalanguage of the truth-conditional theory of meaning and as the language in which instances of the principle of semantic shift are framed. A stipulation of that principle, as applied to sentences, takes the meanings of those sentences as already given. This is in part because, in order to derive from that principle an understanding, whether complete or partial, of the meaning of the word 'true', it is not enough to know, for example, that the statement:

> The statement 'Swans sing before they die' is true if and only if swans sing before they die.

holds good: one must grasp and accept the proposition expressed by that statement, which is to say that one must *understand* the statement. But since the sentence 'Swans sing before they die' is not only referred to, but occurs as a clause, in the statement, this means that one must understand that sentence. But that is not all. In order to recognize a biconditional statement *as* an instance of the principle of semantic shift, you must be able to recognize that the proposition expressed in one clause is the same as that expressed by the statement referred to in the other; you must therefore understand, not only the former clause, but the statement referred to in the other clause. It may be objected that it is enough to notice that the same form of words is used in one clause and mentioned in the other (with due allowance for changes of pronouns, tenses, and so on). But it is *not* enough.

To adapt an excellent example of the late Elizabeth Anscombe's,[8] the conditional

> If Henry can eat any fish, then the statement 'Henry can eat any fish' is true.

is *not* an instance of the principle of semantic ascent. Henry can eat trout; it does not follow at all that it is true to say, 'Henry can eat any fish'.

Thus, by stipulating the principle of semantic shift, as expressed in English, the minimalist takes as given the meanings of English statements. He is claiming, by means of that stipulation, to give a comprehensive account of the meaning of the word 'true', as applied to statements made in English. He must therefore have found some means, adequate in his eyes, for specifying the meanings of such statements; or else he is simply trading upon his readers' already knowing the English language. It follows that he cannot accept the proposal that the meanings of statements, whether of English or of any other language, are to be specified by laying down the conditions for them to be true, in that sense of 'true' of which the minimalist claims to have given an account. For statements made in English, such a procedure would be grossly circular; and the minimalist has given no account of the meaning of 'true', as applied to statements made in any other language.

But can we not get round the difficulty by means of a stipulation that takes no account of the phenomenon of semantic shift, but nevertheless yields all the instances of the principles governing it? Such a stipulation is that embodied in a Tarskian truth definition, adapted to natural language, with the semantic paradoxes avoided by some suitable means (say, by Kripke's construction). The truth definition itself does not recognize or appeal to the occurrence of semantic shift: a recognition of it takes place only in the surrounding patter stating the alleged condition for the adequacy of the definition. The definition simply yields, for each statement of the language, what is known as a T-sentence, namely, a biconditional incorporating, for that statement, the principle of semantic shift, without acknowledging that that is what each such biconditional is.

A truth definition for a language aims, of course, at defining the predicate 'true', as applied to statements of that language. Like all definitions, it takes the language in which it is expressed, save for the term being defined, as understood; and if that language coincides with or is an extension of the object language, it takes the object language as understood. So, although we have circumvented recognition as such of instances of semantic shift, we still have our first reason for attributing to one who stipulates that such instances are to hold good the assumption of a prior understanding of the language, namely, that the clauses of the T-sentences must be understood if the definition is to be understood.

---

[8] Given in personal conversation.

Davidson is well known for having proposed that a Tarskian truth definition for a natural language be turned on its head to provide a truth-conditional theory of meaning for that language. By "turning it on its head" he had in mind that, instead of taking meaning as given in order to explain the notion of truth, we should take the notion of truth as given and use it to explain meaning. To take the notion of truth as given is to know about it whatever has to be known in order to see a Davidsonian truth theory of this kind as a theory *of meaning*, that is, as specifying the meanings of the words and sentences of the language. The word 'true', as used in the truth theory, could not be replaced by some arbitrary nonsense word: it is not a theoretical term that derives its whole content from its role in the theory. We have to bring to the theory our intuitive understanding of the word 'true'. The theory, as stated in Davidson's classic papers, does not seek to make explicit in what this intuitive understanding consists, or to spell out the connection between truth conditions and meaning.

Now, could the meaning of the word 'true', as used in a Davidsonian theory of meaning, be given in the way that the minimalist favors, namely, by stipulating that all legitimate instances of the principles of semantic shift are to hold good? To avoid difficulties about how such instances are recognized, let us suppose that the stipulation is effected by a Tarskian truth definition. That, according to the minimalist, tells us what 'true' means. We have now to use this term to specify the meanings of the words and sentences of the language. To do this, following Davidson's recipe, we simply repeat the truth definition all over again, with the only difference that we no longer call it a *definition*, but a *theory*, what were clauses of the definition have now become postulates of the theory. The procedure is manifestly ludicrous. A minimalist account of truth rules out a truth-conditional account of meaning.

LECTURE 2: STATEMENTS ABOUT THE PAST

To fulfill its task, a theory of meaning must yield an account of use. The most obvious way in which to fashion such a theory is to take as the central notion something that is itself a feature of use; the choice then lies between a justificationist theory and a pragmatist one. If I was right to argue that nothing hangs upon this choice, because they will come to the same in the end, we may concentrate upon the justificationist theory. What is the conception of truth appropriate to a justificationist theory of meaning? Plainly, it must turn on the notion of our being justified in asserting a statement. It is evident at the outset that the word 'our' must be taken in a collective, not a distributive, sense. In learning language, we learn not only when it is right to say certain things, and what responses certain utterances evoke, but how to respond to the utterances of others; such responses include acting on the truth of what others assert. A child who had learned only when he was right to come out with simple assertoric utterances, such as 'Doggie' when a dog was in sight, would serve as an extension of adults' range of observation, but could not yet be credited with saying that anything was so: he can be credited with that only when he has learned to treat the utterances of others as extending *his* range of observation. It is intrinsic to the use of language that we accept the testimony of others: to believe what we are told is the default response. Language binds us into society.

The intuitionist theory of meaning applies only to mathematical statements, whereas a justificationist theory is intended to apply to the language as a whole. The fundamental difference between the two lies in the fact that, whereas a means of deciding a range of mathematical statements, or any other effective mathematical procedure, if available at all, is permanently available, the opportunity to decide whether or not an empirical statement holds good may be lost: what can be effectively decided now will no longer be effectively decidable next year, nor, perhaps, next week.

All the same, we can gain some insight into the matter from the intuitionistic theory of meaning. Intuitionistic semantics specifies the meaning of each logical constant by laying down what is to count as a proof of a statement of which it is the principal operator. But only a little reflection is needed to show that 'proof' in this context means something narrower than an intuitively valid deductive argument. A proof of a statement of the form '$A \rightarrow B$' is specified to be an operation that can be recognized as taking any proof of A into a proof of B. An argument employing modus ponens is, in that respect, intuitively valid by anyone's standards: so, if what was here meant by 'proof' was anything that would count as an intuitively valid argument, anything whatever would serve, according to the specification of the meaning of the connective '$\rightarrow$', as a proof of '$A \rightarrow B$', since by modus ponens we could take it and any proof of A into a proof of B. The same applies to the rule of universal instantiation. A proof of a universally quantified statement '$\forall x\, A(x)$' is specified to be an operation that can be recognized as taking any element of the domain into a proof of the corresponding instance: plainly, the proof of the instance cannot be allowed to appeal to universal instantiation, on pain of trivializing the specification of the meaning of $\forall$. Clearly, by 'proof' in these contexts must be meant an especially direct kind of proof, which may naturally be called a canonical proof.

Plainly, however, if a canonical proof were to be demanded of anyone who purported to have proved a statement, it would be impossible for him to appeal to theorems previously proved. If a previous theorem was a universal statement, he would be unable to appeal to it to justify invoking an instance of it in the course of his proof: he would have actually to apply the operation embodied in the proof of the earlier theorem to obtain a canonical proof of that instance. This would not only greatly increase the length of his proof, perhaps beyond manageability; it would stultify the entire practice of storing up mathematical information. As Frege remarked, theorems and their proofs store up deductive routines which do not have to be gone over again, but can simply be appealed to. That is why elimination rules that cannot appear in the body of a canonical proof, such as modus ponens and universal instantiation, can be employed in an intuitively valid demonstration.

The property that we seek to preserve in an intuitively valid argument, which is also the property required to justify asserting a mathematical statement, is therefore not possession of a canonical proof: it is having an effective means—a means effective in principle, even if too laborious for practice—of constructing a canonical proof. In mathematics, a constructive proof is one that shows how to prove the conclusion, given that we can prove the premises. Thus, Leonhard Euler's solution of the Königberg bridge problem shows how to find, from an itinerary involving the crossing of every bridge, a bridge that was crossed twice. In an empirical case in which a walker has been observed to cross every bridge, we may infer that he has crossed some bridge two or more times. But it does not follow that, if we have verified the premise, we can verify the conclusion. We can easily conceive of observers stationed at each bridge, each of whom leaves his post as soon as he sees the walker crossing that bridge but reports only later without giving the time of crossing: we have then no means of identifying a bridge he has crossed twice. In the empirical case, therefore, the conclusion of a valid argument from premises that have been verified is guaranteed to be such that it *could have been* verified, not that we can any longer verify it. If a mathematical statement could have been verified, then it can still be verified, but that is not so for an empirical statement. We ought not to demand a stricter standard for valid deductive arguments about empirical matters than we do about mathematical matters: the difference between the two cases requires us to place a lighter burden on the conclusion for it to be true and hence to be believed.

This conclusion—this conception of the notion of truth appropriate to a justificationist account of meaning—must come as a relief to anyone attracted to such an account of meaning yet troubled about the reality of the past. If that account of meaning demanded that we allow as true only those statements about the past supported by present memories and present evidence, then large tracts of the past would continually vanish as all traces of them dissipate. We should be committed to a metaphysical conception according to which nothing exists but the present: the past would be a mere construct out of whatever in the present we treat as being traces of it. We could not so much as think of a statement about the past as having once been true, though now devoid of truth value, save in terms of present evidence that evidence for its truth once existed. This conception, though not incoherent, is repugnant: we cannot lightly shake off the conviction that what makes a statement about the past true, if it is true, is independent of whether there is *now* any ground that we have or could discover for asserting it. But if the truth of a proposition consists of its being the case that someone suitably placed *could have* verified it, or have found a cogent ground for asserting it, then our conviction is vindicated. On this conception of truth, a statement about the past could be true if someone at the relevant time could have verified it, even though all reason for asserting it may have blown away.

For convenience, I shall use the term 'verify' to cover what falls short of verifying a statement in the strict sense; I shall treat it as covering also possession of grounds sufficiently compelling to warrant asserting the statement. Such grounds would *justify* the assertion, which is why I am speaking of a "justificationist" rather than a "verificationist" account of meaning. It is of interest to seek to make this notion precise; but that is not what concerns us in the immediate context.

I confess to being in just the position of the thinker inclined to a justificationist theory of meaning but troubled about the reality of the past. I am indeed attracted to a justificationist account of meaning; at the same time, I have long been concerned to reconcile the reality of the past with that account. Many years ago, I wrote and published an article[1] to explore the issue. When I began to write it, I had hoped to arrive at one or other conclusion: that antirealism about the past was a benign and acceptable view; or that it was incoherent, and that its incoherence would expose a fallacy in the argument for a justificationist theory of meaning. In the end, the conclusion that I reached was the most disappointing possible. Antirealism about the past was *not* incoherent; but it was not believable, either. I have been perplexed by this matter ever since.

It is not enough to adopt the formula "what could have been verified" as encapsulating a satisfactory justificationist conception of truth. We have not vindicated such a conception. We have not shown how, if meaning is to be explained in accordance with a justificationist theory, a speaker may be supposed to acquire a tacit grasp of the notion of truth so conceived. We have merely argued that, if a justificationist is to allow the validity of constructive deductive arguments—those which provide a means of verifying the conclusion, given a sufficiently detailed verification of the premises—he must admit the capacity of a statement to have been verified as a sufficient condition of its truth. The verification of the premises may not have been sufficiently detailed. Even though the argument may convince us that it must have been possible to verify the conclusion, the time for doing so may have passed. If that conclusion relates to the distant, or often even to the fairly recent, past, it will almost certainly have done so. When a constructively valid argument relates to the past, it is possible that we should have a record of its premises having been quite conclusively verified, but that a verification of its conclusion is now forever beyond our reach. But all this shows is that an adherent of a justificationist theory of meaning must choose either of two positions: (1) to deny that constructive arguments, when applied to empirical situations, and particularly to those which lie in the past, are always valid;

[1] "The Reality of the Past," *Proceedings of the Aristotelian Society*, n.s., LXIX (1969): 239–58, reprinted in *Truth and Other Enigmas* (Cambridge: Harvard, 1978), pp. 358–74.

and (2) to admit that the mere possibility of its having been verified—in an appropriate sense of 'possibility'—is a sufficient condition for the truth of a statement. It does *not* establish that truth is to be *identified* with the possibility of being or having been verified; it does not show that that is the conception of truth at which a speaker will arrive if a justificationist theory of meaning is correct.

The justificationist therefore needs to engage in a good deal more argument before he is entitled to claim that the dilemma about the reality of the past has been resolved. Above all, he needs to supply a plausible account of our understanding of statements about the past, and of how we come by that understanding, for it was an application to such statements of the general form of justificationist—or antirealist—considerations about the acquisition of understanding that drove him in the first place to regard an understanding of the past tense as consisting in a grasp of what would *now* warrant or justify assertions concerning the past.

How do we come by an understanding of statements about what is spatially remote? A child first learns to grasp the three-dimensional disposition of the contents of a small space, such as a room in a house or a small garden. He then learns the layout of some adjacent such spaces—first the rooms in a house, and next the neighboring houses to his own. He then progresses to the construction of a mental map of his near surroundings—the streets and parks or the fields and woods to which he can walk. He is learning the conception of things—houses, trees, pools, rivers—as being disposed in three-dimensional space, and of movable things—animals, people, cars, clouds, birds, airplanes—as moving about in that space. At each of these stages, it is given to him to connect the concept of a place with that of his own movement, that is, of his going to that place. He comes to grasp the notion of distance, and to understand the relation of the distance between two things to the time that it takes to get from one to the other; the conceptual connection between space and time is fundamental indeed. This connection is not completely tight: there are places which he can observe but cannot go to, such as the air in which birds and kites fly, the sky in which the clouds move: but—although he may need to be corrected about rainbows—it is crucial that he understands the birds and clouds as objects of which it can be asked how far away they are, and are hence in the same space as we are. Obviously, the spatial range of which he grows aware progressively expands: he is taken by car or train to distant places, and later taught first the use of a map and then of a globe. To understand a globe, he has of course to form a conception of the Earth as a whole and as roughly spherical in shape and having an interior. As the range of places of which the child can conceive grows greater, he retains his primal understanding that it makes sense to ask, of any place, what is happening there now. At a late stage, he comes to understand that the heavenly bodies—the moon, the sun, and the stars—are material bodies too, located at great distances which are still expressible in terms of the miles by which we measure terrestrial ones.

What is important in this process of coming to understand reference to places more or less distant from that which we occupy is that it requires the mental construction of a grid—a system of co-ordinates by reference to which different places can be specified: for coming to understand reference to different times likewise requires the mental construction of a grid, though in this case only a linear one. The process of learning to comprehend reference to different places at first sight wars only to a minor extent with a simple justificationist account of understanding. At an early stage of a child's intellectual growth, it is natural for him to interpret statements about how things are at some other place in terms of going to that place and making the required observations; one may not want to do this, but the possibility of doing so gives the content of statements of this kind. The interpretation will not do when it is impossible to reach the place in question sufficiently quickly to make the necessary observations. We have then to conceive of going to that place and determining whether the statement held good there at the time at which it was made. This, of course, raises the problem of our understanding of statements about the past; so there is nothing special about statements about remote places which causes embarrassment to the justificationist.

A child begins to learn the use of the future tense from expressions of intention, in particular both threats and promises, made by his parents or other adults: he comes to understand the succession of days, and so to form a grid that stretches forward. At the same time, he starts to learn the use of the past tense, and so to extend the grid backward. He is given the past tense when he comes out with avowals of his own memories, and when adults evoke his memories by recalling events in their recent joint past. Realist accounts of thought and linguistic understanding routinely invoke the conception of knowing what it is for such-and-such a state of affairs to obtain or for such-and-such a type of event to take place. This becomes problematic when the state of affairs or type of event is one that cannot be directly recognized as obtaining or taking place; but this is not a problem to do with the past tense as such, and we may assume without loss of generality that the statements made to the child about the past, or those of his utterances which are interpreted as tantamount to such statements, all concern states of affairs or types of event that can be directly recognized when they presently take place or have just taken place. With this reservation, we have no difficulty with the notion of knowing what it is for such-and-such a state of affairs to obtain or for such-and-such a type of event to take place: to know this is to be able to recognize the state of affairs as obtaining or the event as occurring when suitably placed to observe it. Nor is there as yet anything problematic about the conception the child acquires of what renders true a statement about the future or one about the past. To establish the truth of a statement about the future, it is necessary to wait for the required length of time, and then to observe whether the state of affairs or event foretold obtains or occurs. To establish that of a statement about the past, one must rely on one's own memory or that of others.

So far, then, it appears that the child's understanding is perfectly in accord with a justificationist account of meaning. His understanding of statements about the past must indeed be enlarged so as to connect them, not only with memories, but also with traces of past occurrences: he must learn to say things like, 'He has left the book on the piano', 'She must have switched the heating on when she went downstairs', 'The cat seems to have killed a bird on the lawn', and the like; but, of course, this enlargement remains fully consonant with the justificationist account of the meaning of statements about the past. But caution is required in acknowledging this harmony between a child's understanding at an early stage of his conceptual development and the account of understanding derived from a justificationist account of meaning.

A gap already begins to appear between what establishes a statement as true and what it says holds good. Nothing analogous can be found in the intuitionistic account of the understanding of a mathematical statement. You understand such a statement, according to the intuitionists, if you are able to recognize a proof of it when presented with it. It can be established as true by obtaining such a proof; and what it says is that there is such a proof. The proof is simultaneously what is required to establish the statement as true and that of which an assertion of it warrants the existence. This is equally so for many empirical statements. For instance, the canonical way of establishing the truth of such a statement as 'There are seventeen apricots in that bowl' is to count the apricots and find that they come to seventeen. And this is just what the statement says, namely, that if you count the apricots you will find that they come to seventeen.

A gap between the two becomes apparent, however, with empirical statements about how things are in another place. The most direct way available to us of verifying such a statement is to go to that place and observe that that is how things are there, and a grasp of this is integral to the child's coming to understand the meaning of such a statement. But that that is what you would observe if you went there is not what the statement *says*; a child who thought that would not yet have come even to the fundamental stage of understanding statements about other places. We noted that it is integral to the process of learning language that an infant learns not only to give voice to what he himself has observed (as well as to his wants), but that he learns also to grasp the utterances of others as informing him about the world (as well as telling him what he is to do). In the process, he has to make the different pieces of information he receives, from his own observation and from what others tell him, consistent with one another, which will often require that he modify them: they have to cohere into a single intelligible picture of the world. The formation of the mental spatiotemporal grid that his understanding of statements about other places and about past and future times requires him to construct involves that he comprehend such statements as locating states of affairs and events at points on that spatiotemporal grid. His grasp of what states of affairs and types of event those are derives from the ability he has acquired to recognize them when he observes them. When they are said to obtain or to have occurred elsewhere, he knows that he would have to go to that place in order to observe them, and thus to establish the truth of what is said. But that is not how he takes what he is told to be *saying*. What it says is that at that particular location on the spatial map is something of a kind he can recognize when he himself is at the right location.

The distinction just drawn between what is needed to establish a statement as true, and what the statement says, needs careful scrutiny. If we adopt it, we take a very decided lurch in the direction of a truth-conditional account of meaning, and thus of a realist account of our thought and language. In drawing the distinction, I have deliberately avoided invoking the notion of meaning. There is a strong temptation to say such things as 'The statement "A rowing race is taking place at place *q*" does not *mean* that, if you were to go to *q*, you would observe a rowing race'; and 'The statement "It rained here yesterday" does not *mean* that those who were here yesterday remember its raining': but appeal to the notion of meaning in making the point begs the question. The justificationist theory of meaning takes the meaning of a form of statement to be given by what is needed to establish it as true. A radical justificationist regards this as being that in virtue of which the statement is true, if it *is* true. That is why the antirealist about the past, in my paper of long ago on "The Reality of the Past" (*ibid.*), ended by holding that there only ever *is* what exists in the present, and hence that a statement about the past, if true at all, can be true only in virtue of what *now* exists or holds good. He cannot allow that, while such a statement can now be established, if at all, only by what lies in the present, what it says is not about the present. We normally suppose that a statement about the past, if true, is rendered true by what lies in the past: hence, although in one sense the past is no more, it must in a definite sense still possess a shadowy existence, since otherwise there would be nothing to render statements about the past true or false. But if someone is genuinely an antirealist about the past, he thinks that the past, as the past, retains no existence whatever: the past survives only in its present traces, including memories. That is why, in distinguishing between what can establish a statement about the past as true and what it is that that statement says, we are repudiating antirealism about the past.

The antirealist's idea that the past is genuinely no more and does not possess even the shadowy existence needed for it to render statements about it true or false, has a certain attraction, even though it conflicts with our natural realism about the past. But the idea that other places do not of themselves exist, but subsist only in their effects on where we are now, has no attraction whatever; this is why I discussed the child's learning to understand statements about other places before talking about his learning to understand statements about other times. At least as far as places on or near the Earth's surface are concerned—or in virtue of space travel, we should perhaps say within the inner solar system—to adopt such an idea would make us solipsists: it would run counter to the principle that we each must accept as true what any of us has established as true, that verification is not an individual but a collective activity.

As already remarked, according to the justificationist theory of meaning, the meaning of a form of statement is constituted by what is needed to establish it as true. So does the gap that there may be between what is needed to establish a statement as true and what the statement says demand that we add a further component to its account of meaning: to allow that, as well as what is needed to establish a statement as true, there is also what the statement is understood as saying? Well, *why* do we attribute to the child an awareness that what the statement 'Your sister must now be sitting down to her breakfast' says is not that, if he were to go to where his sister is, he would observe her having her breakfast, even though he knows that that would be the surest way of making certain that the statement was true? We do so because we credit the child with a consciousness that this would not be the most *direct* way of verifying the statement. Just as with mathematical statements, understood intuitionistically, so we need a distinction between the canonical, or most direct, way of establishing an empirical statement as true and legitimate, but indirect, means of doing so. The canonical way of doing so will always be that which corresponds to the composition of the statement and the way the meanings of the words that make it up are given. Thus a casual glance may be all that is needed to confirm the statement 'It is either raining or hailing', but this is not yet the canonical means of doing so: in virtue of the meaning of the connective 'or', the canonical way to do so must be to establish either that it is raining or that it is hailing. Consider the statement, 'There are seventeen apricots in that bowl'. Virtually anyone, asked to characterize the meaning of that statement, and certainly anyone disposed to explain the meaning of a statement in terms of how we find out that it is true, would talk in this connection about counting: counting the apricots is the direct or canonical method of establishing its truth. But suppose I have only just put the apricots in the bowl, having just bought them and having bought nothing else. Then, if I find I have spent $2.89, I can conclude that there are seventeen apricots in the bowl, since 289 is the square of seventeen and I know that they did not cost $2.89 each, and I know that they cost more than 1 cent apiece. This is a perfectly cogent piece of reasoning, but it is not a direct or canonical way of establishing the truth of my conclusion.

We attribute to the child a grasp, doubtless inchoate, of the difference between a direct and an indirect way of making sure that something is so. Why should we credit him with regarding the procedure of taking a journey to a place and observing what is there as no more than an indirect way of establishing the truth of a statement about how things are there? Well, what is the alternative? Let us grant to the child an ability to tell where he is—by no means in all cases an easy thing to do, as the child will know if he has ever lost his way. But we shall not attribute to him an understanding of what, in general, it means to say that an event of a certain kind is happening at such-and-such a place if all he knows is how to determine when he is in that place and how to observe whether an event of that kind is happening there. He must indeed know that, if he is to be credited with an understanding of statements of that type. But he must also understand that already to be in the place referred to, and to observe what is taking place there, is the only *direct* way to verify the statement, that it may be true even though such a direct verification of it is not available and that its truth is capable of being indirectly established. In other words, he must have advanced well beyond the point of grasping only statements of the most elementary kind: those, namely, which we are able to verify by immediate observations, and which are falsified by our failing to make such observations. 'It is now raining (here)' is such an elementary statement. To verify it, you need only look about you or hold out your hand, or, if you are indoors, look out of the window or front door: if you do not then see or feel that it is raining, you know that it is *not* raining. There is seldom any role for the indirect verification of such elementary statements to play. But, although a statement that a certain event is occurring in a certain place can be directly verified by the child's establishing that he is in that place and observing the occurrence of the event, it is not falsified by his establishing that he is *not* in that place. The statement does not mean that he is in that place and can observe the occurrence of the event: rather, it has the conditional meaning that, if he were in that place, he would be able to observe the event's occurrence. It therefore follows that any evidence the child may have for the truth of the statement—including his traveling to that place and observing the occurrence of the event—must be categorized as indirect.

This might be objected to, on the ground that the child's understanding of reference to different places is founded upon his grasp of how it is possible—when it *is* possible—to get to them. His grasp of the meaning of a statement about what is happening at a distant place therefore incorporates the conception of getting to that place and observing how things are there. To some extent this is correct. Even an adult may have no better grasp of where some street in his town, or some nearby town, is located than a knowledge of a route to it. But this is only a partial grasp of where that street or town is. For one whose entire understanding of the meanings of expressions referring to that street or town was constituted by a knowledge of a route to it, it would be unintelligible to speak of a different or better route to that street or town: a street or town reached by a different route could not be for him the *same* street or town. That is why a full understanding of references to different places, and hence of statements about what is happening elsewhere, demands the conception—however rudimentary—of a spatial grid upon which a map may be constructed.

Thus, reflection upon our understanding of statements about what states of affairs obtain, or what events are occurring, in other places forces the justificationist a certain distance in the direction of realism.

The grasp of such a statement falls into two parts: one is an understanding of what it is for a state of affairs of the type in question to obtain or an event of the type in question to occur; the other is our knowledge of how to locate it on the grid that serves to particularize the place referred to. This bears a resemblance to the realist's idea that our understanding of a statement ascribing some property to an object is constituted by a knowledge of what it is for an arbitrary object to have that property, together with an understanding of reference to that object. The justificationist may still explain the former constituent—our grasp of the relevant type of state of affairs or of event—in terms of our learned ability to recognize that state of affairs as obtaining, or that event as taking place, here and now. But he must relinquish any idea he had that our understanding of statements of this kind has only one ingredient, rather than two: that is, that it is constituted by a knowledge of what affords a ground for accepting each such statement as a whole. It indeed remains, for him, integral to that understanding that we are able to recognize evidence for accepting it: moreover, that we know that the most nearly direct grounds for doing so are supplied by going to the place in question and observing how things are, or what is going on, there. But it is also essential to a sound account of our understanding of such statements that any evidence to be gathered by those who are not now in the place referred to must count as indirect.

The child's conception of the spatial grid will, of course, undergo refinement as he matures. He will first acquire the knowledge that the Earth approximates to a sphere in shape; then he will learn the rudiments of astronomy. Despite the boastfulness of contemporary scientists, it took the human race a long time to grasp the rudiments of astronomy. I should be surprised to be told that, if civilization lasts that long, which I doubt, in five centuries from now people will not regard the conceptions of present-day cosmologists as being as far from the truth as we regard the Ptolemaic system.

Of course, many states of affairs and many occurrences cannot be recognized as obtaining or taking place by mere observation. In my discussion of statements about other places, I in effect confined myself to those which describe states of affairs and events that can be so recognized as obtaining or as happening in some other place. This was simply to isolate the matter of our understanding of reference to other places: in order to isolate it, we must assume it as already known what it is to conceive of the state of affairs or event in question as obtaining or taking place, and the most straightforward way to do this is to assume that it is one that can be recognized by simple observation.

LECTURE 3: THE METAPHYSICS OF TIME

On a constructive interpretation of meaning, a mathematical statement is true if we possess a means, effective in principle, of constructing a canonical proof of it. Now, if we possess a procedure for finding a proof of one or other clause of a disjunctive statement, we do not need to know of which of them it would, if applied, yield a proof. Having such a procedure, we may assert the disjunctive statement as true. The example of vague statements shows that it is not in itself abhorrent to allow a disjunctive statement to be true even though neither disjunct is true; but in the mathematical case we may hold that, for any true disjunctive statement, one of the disjuncts must be true, even if we do not know which. That disjunct would be true in virtue of our possessing a procedure that *would*, if applied, yield a proof of it, even though we may not know it would. So the condition for the truth of a mathematical statement, under this conception, would be a hypothetical one.

What, then, should a justificationist require for a statement about what is happening at a given place to be true? In speaking of a child's acquiring an understanding of statements referring to particular places, we were concerned with his coming to understand statements about places other than where he was. But he does not acquire a solipsistic conception of linguistic meaning: he is taught to accept what others tell him on the basis of their observations. By means of our language we distribute information among ourselves, as well as from past observers to later generations. So there is direct evidence for the truth of a statement that something is occurring at a certain place provided that someone is observing it there. The observer need not be at the very place in question, but only sufficiently close in time and space to observe an event of that kind; an event taking place at an astronomical distance from any observer must be one occurring when the light signal was emitted. Observation need not be confined to that effected by the unaided senses, but may also comprise observation by means of instruments; it may be made by any creature with whom we are able to communicate.

Realists think of our understanding of all statements whatever as consisting in the knowledge of what it is for them to be true. They usually simply help themselves to this notion: they show little interest in explaining what our conception of what it is for statements of different kinds to be true consists in, let alone how we come by such conceptions. My discussion of a child's coming to understand statements about other places was intended to explore how a justificationist is forced, if he aspires to give a credible explanation of our understanding of our language, to take a step in the direction of realism. For the notion of what a statement says, as contrasted with how it can be established, approaches the realist conception of truth: it has to do with that in virtue of which the statement is true, if it is. For the justificationist, it has to do with direct evidence for the truth of the statement; but the notion plays very different roles in the empirical and mathematical cases. To have indirect grounds for the truth of a mathematical theorem is to have an effective means, in principle, of constructing a canonical proof: but indirect evidence for the truth of an empirical statement supplies no means of obtaining direct evidence. How much must the justificationist concede to realism without surrendering his fundamental conception of meaning?

What criterion should the justificationist have for the truth of statements about unobserved events? The natural criterion is one expressed by a conditional: if someone were to have been at hand at the relevant time, he could have observed an event of that kind. This is certainly enough to vindicate indirect ways of finding out what is happening in places beyond the scope of any observer; but it is not enough to vindicate the principle of bivalence for such statements. We are certainly not in general entitled to assume that every conditional with an unfulfilled antecedent determinately either holds good or fails: the most resolute realist would not make such an assumption, because the consequent might depend upon some factor not mentioned in the antecedent. The assumption *is* plausible for a conditional of the form, 'If such-and-such a mathematical decision procedure were applied, it would yield an affirmative result'. It is plausible because the outcome of the decision procedure is unaffected, at each step, by external factors, but is internally determined. That is not so with the empirical procedure of making observations: the outcome of this procedure *is* determined by factors external to it. To assume that there is always a definite truth about what could have been observed by a sufficiently near observer is to assume that the world is determinate independently of our experience; and this is a realist assumption, indefensible from justificationist premises.

If *A* is a present-tensed statement decidable by observation, then we must allow the truth of '*A* or not *A*'. The principle that logical laws applicable to statements in the present tense must carry over to corresponding statements in the past tense therefore implies that, where *A** is a past-tense transform of *A*, we must also allow the truth of 'A^* or not A^*'. This could therefore be made a ground of objection to our rejecting bivalence for observational statements in the past tense. The objection would be unsound. If no observer was present, we do not need to assume that either 'If one had been present, he would have verified that A^*' or 'If one had been present, he would have verified that not A^*' is determinately true; it is sufficient that 'If one had been present, he would have verified either that A^* or that not A^*' is true. Any conclusion we were to draw from 'A^* or not A^*' would still be under the hypothesis of the presence of an observer.

A justificationist semantics allows for there being gaps in reality: questions to which there is no answer one way or another. Gaps in reality are like gaps in a fictional world. A question about the color of Horatio's eyes may *have* no answer: there is no truth of the matter. There will be many unobserved events for whose occurrence we have indirect grounds; but, when we do not, we are not entitled on justificationist principles to assume that there must always be some true answer to the question whether they occurred or not.

As I have often insisted, direct evidence for the truth of an empirical statement will not in general consist of bare observations, but may comprise an inferential component. At one end of a spectrum lie utterances that may serve as reports of observation, at the other mathematical theorems, established by pure deductive reasoning. Most statements we make lie between these extremes: they may be most directly established by reasoning from premises based upon observation. Among these belong statements about physical conditions that could not be observed—conditions in the interior of the sun, in a supernova, within a black hole; perhaps also all statements about conditions before there were any observers. We arrive at judgments about such matters by reasoning, in accordance with our physical theories, from observations more or less remote from our conclusions: we could not arrive at them in any other way. The contrast here is not metaphysical: it is merely between what can be established by simple observation and what it requires inference to arrive at. As I have said more than once in these lectures, I have been assuming the state of affairs which we are asserting to have obtained in the past to be one directly observable when holding good here and now solely in order to isolate the problem what effect the past tense has upon the meaning of the statement. Here we have a genuinely metaphysical disagreement. I have been arguing that it is consistent with a justificationist theory of meaning to take direct evidence for the truth of such a statement to consist in observations made from a suitable distance at the relevant past time; indeed, that it is essential so to take it if the theory is to treat evidence, not as what an individual subject possesses, but as what is in principle available to us collectively. We want a theory of meaning for our common language, not for an idiolect; inasmuch as all human languages are essentially intertranslatable, there is a single language common to us all. The antirealist about the past holds that evidence for a statement about a past event can consist only in present memories or in presently observable traces of that event, together with any reasoning required to derive the event from those traces. This is a metaphysical controversy; it is this I am attempting to explore.

Statements about other places were, of course, intended as an analogue to statements about other times. Here, too, the child must acquire the conception of a grid, indeed, a spatiotemporal grid. But can we attribute to the child a recognition that present evidence for the truth of a statement concerning another time is indirect? Must he, to have acquired an understanding of the past and future tenses, grasp that what a statement couched in either of those tenses *says* is not that such evidence exists, but, rather, that that is how things were or will be at the indicated position on the temporal grid? Do we have a true analogy with his understanding of statements about other places?

If the analogy is a true one, then what is *said* by a statement in a tense other than the present is that an event occurred or will occur at the time and place it refers to. Evidence available at a different time may render the statement probable or even establish its truth, but it can only be indirect, since its being available is not what the statement *says*: it is not what renders the statement true. Direct evidence for the truth of the statement can consist only in observations made at whatever time the event could be observed by an observer placed sufficiently close to observe it.

One who denies the reality of the past will object. Statements about the past, he says, ought not to be conceived on the analogy of statements about distant places. We can, in some cases, verify a statement about a distant place by going to that place and observing how things are there: but the whole point concerning statements about the past is that we *cannot* in principle locate ourselves at the times to which they refer; the past, as the past, is inaccessible. That, he argues, is why we must consider the only direct way of establishing the truth of a statement in the past tense—the only way in principle available to us—to be on the basis of its present traces: the memories of those who witnessed how things were at that time, the reports that have come down to us from past observers, the causal consequences of what happened then. The past is not a region we can travel to, he declares: it persists only in its present traces. The past has no other existence: the past, *as* the past, is no more. All that now exists is the present: only what now exists can render any statement true or false.

This objector has plunged into tumultuous metaphysical waters. The mistake he makes is to explain meaning in terms of an individual's means of verifying the truth of a statement; that is why he takes the direct way of establishing the truth of a statement about a distant place as consisting in a journey to that place, followed by suitable observations, rather than as an observation made by an observer in that place. That is not the proper form for a justificationist theory of meaning, which is concerned with how *we*, collectively, can establish a statement as true. If, as we have been assuming for simplicity, the statement is of a kind that can be established by observation, then that is the direct way of so establishing it. An observer suitably placed may make such observations; and he may then transmit the information that he has so acquired to others. Someone else to whom this information has been passed on will not himself have made the relevant observations. His personal possession of the information is founded upon testimony; but its original source was founded upon observation. That is one of the central functions of language, namely, to spread knowledge to the community at large. For that reason, a theory of meaning for our language—that is, ultimately, for all intertranslatable languages—cannot be based upon an individual's use of the language to express the knowledge that he personally has: it must be based upon our common use of language to express and transmit knowledge accessible to us all. Indirect evidence for the truth of a statement about a place where no observer was present must show that a suitably located observer could have made observations giving direct grounds for its truth: that is how indirect evidence must be related to direct evidence.

In the same way, direct evidence for the truth of a statement about the past, of the kind that could be a report of observation when made in the present tense, takes the form of an observation made at a time when that observation was possible. If someone still living did make such an observation, he may of course report it to others on the strength of his memory of doing so. Not only the living may report their observations. If we have a written record of what was observed by a witness who is now dead, or an account of his observations which has been passed down a chain of informants, we are in as good a position to know by testimony that the past event was observed as if we were told by a living observer. Dying does not deprive anyone of the status either of an observer or of an informant: the dead remain members of the community with whose collective and imperfectly shared knowledge we have been concerned. Admittedly, most of their messages have been obliterated by time; but many of the dead still communicate with us through their words that have been preserved, their writings, their works of art, and their scientific and philosophical theories. For all the messages that have been lost, it remains that statements about the past must count as having been directly established, and therefore as true, if someone observed them to be true at the appropriate past time.

A justificationist must accept this account of how statements about the past are to be understood because it is incontrovertibly how we do understand them: antirealism about the past does not faithfully represent the manner in which we, in fact, understand the past tense. In accepting it, he will have substantially modified his doctrine. If, for every statement in our language, there were an effective means of deciding its truth or falsity, there would be no practical difference between a justificationist and a truth-conditional explanation of meaning. It is largely because we can frame statements that are not in this sense decidable that such a difference exists and has a substantial effect both on the logical laws accepted by the adherents of the two types of meaning theory and on their metaphysical conceptions of reality. Now, if we assume that statements of the most primitive type are in principle effectively decidable, the question arises what components of sentence formation import undecidability into our language: What makes it possible for us to frame undecidable statements? I once wrote[1] that there are three linguistic devices that have this effect: quantification over infinite totalities, as expressed by such a word as 'never'; the subjunctive conditional form; and the past tense. I have been arguing that the past tense is out of place in this list; its inclusion is due to thinking that direct evidence for what happens at some place and time consists in traveling to that region of space-time. Adopting the revised account of statements about the past will modify the justificationist's theory of meaning in a realist direction, but will not demand that he repudiate his general principles: he will still hold that a statement about the past can be true only in virtue of an actual or possible direct verification of it.

Observation is ordinarily understood as revealing what would have been so if the observation had not been made. This is a sophisticated thought, which ought not to be attributed to a child who has been taught to say how things are by looking and seeing, or feeling or listening. Rather, it is in part by learning these skills that the child forms the conception of an objective environment independent of his will and of its being observed by him or others: a conception that rounds out the infantile expectations of repeatable surroundings and familiar experiences he has come to have at the earliest stage of life. He learns what to expect to be still the same when he views it afresh, and what to expect to have altered, in accordance with some process that was then under way; he learns to seek an explanation when such explanations are falsified, in terms of something that went on in the interim. It is the standard practice to treat observation as disclosing some state of affairs, rather than as contributing to creating it: that is the practice in which the child learns to engage. Small children do not need to acquire the notion of an objective feature of the world, as they observe it: rather, they need to be instructed which features of their experience are subjective. It is therefore a mistake to argue that a conception of reality as existing independently of being observed must be prior to and inform the observational practice that we learn: it is by learning that practice that we acquire such a conception. That does not mean that our grasp of what it is for this or that observable state of affairs to obtain does not rest upon our learned ability to recognize it as obtaining when we observe it. Quite the contrary: it is precisely our ability to do this that constitutes our primitive grasp of what it is for it to obtain. It is only at a much later stage, as adults, that we extend our childhood distinction between subjective and objective into one between how things are in themselves and how human beings perceive them, a distinction that is far from being philosophically straightforward.

---

[1] "What Is a Theory of Meaning? (II)," in Gareth Evans and John McDowell, eds., *Truth and Meaning: Essays in Semantics* (New York: Oxford, 1976), pp. 67–137; reprinted in *The Seas of Language* (New York: Oxford, 1993), see pp. 46, 60.

When, as in quantum mechanics, there is a theoretical ground for thinking that observation disturbs what is observed, it becomes problematic whether there is any reality independent of observation; if there is, it of course cannot be explained in terms of what could be observed. When the ground is a merely practical one, as in observing a deer or bird liable to run or fly away if we try to observe it, there is no reason to reject an explanation of a statement about that creature in terms of what might be observed: it is just that observation is difficult. In such a case, we can perfectly well conceive of observations made in such a way as to inhibit the creature's flight—for example, by means of effective camouflage or with a telescope. We do not need ourselves to be able to carry out such an operation; it is enough that we can conceive of it.

At first sight, metaphysical views about time may be classified into four categories, according to whether past and future are considered real: one may be real but not the other, both may be real, or neither. What is meant by asking whether the past or the future is real? Some believe that the future is not real, in the sense that there is nothing in virtue of which a statement in the future tense can be true. That is, no such statement can be true or false *now*, when it is made and relates to how things are to be later; it can become true or false at the time to which it relates. A variation is the view that it is an illusion to think that a statement in the future tense is, now or later, rendered true or false by what happens later: the only intelligible use of the future tense is that in which a future-tense statement relates to what is available in the present. We do indeed have such a use of the future tense; it is exemplified by a notice of the form 'The marriage between *X* and *Y* will not now take place', or the statement 'They *were* going to be married, but they are not going to any longer'. On this second view, that is the *only* use of the future tense we can have, and all our statements in that tense must be so interpreted.

There is a disanalogy between the past and future in this regard. No denial of the reality of the past takes the form of holding that no statements in the past tense are true at all. No one says that such a statement may have been true or false at the time to which it relates, but now is no longer either. The reason is evident. To deny that a statement in the future tense can be true or false in advance of the time to which it relates, but becomes true or false at that time, is not blatantly inconsistent with our use of the future tense; after all, we must allow a statement in that tense a degree of probability. But to deny that a statement in the past tense can be true or false after the time to which it relates deprives our use of the past tense of all meaning. We could not even grant it a degree of probability. The probability of a statement in the future tense is that of its proving to be, or becoming, true at the relevant time, and this we are accustomed to from the practice of betting. But the probability of a statement in the past tense can only be the probability of its being now true; if it has no truth value, it can have no probability.

To disbelieve in the reality of the past is, rather, to hold that a statement in the past tense, to be true, must be true in virtue of something lying in the present. But this view is not in full analogy with the second type of denial of the reality of the future. It is not proposed that we should say things like, 'She now married Edward in 1825, but she probably will not have done so next year'. We are bound to acknowledge that the evidence available at a given time for some past event may point in a different direction from that available at an earlier time; but we should not express this as that event's having been different at that earlier time, but simply as the evidence's having then been misleading. To say that the marriage between *X* and *Y* will not now take place is not to say that the future has changed, even if we think that this exemplifies the only future tense that we have; but to allow that a statement about the past was true, but now is false, would be to say that the past had changed. That use of the future tense under which a statement is rendered true or false by what lies in the present may be called the present future: we have no corresponding use of the past tense. Why this difference? The reason lies in our understanding of the direction of causality. We can have evidence for a later or for an earlier event; but only a later event can be prevented. To make a statement in the present future is to say that such-and-such will happen if things go on as they are at present; but something may occur to prevent their doing so. If Jonah's threat that Niniveh was going to be destroyed is understood as having been in the present future, it was true when made; that the destruction did not take place was due to the repentance of the Ninivites, unforeseen by Jonah. There may be cases in which no prevention is possible; but that grounds for a prediction are, in general, subject to that possibility is the reason for our having a present-future use of the future tense. When we have evidence for the occurrence of a past event that no one remembers witnessing, we know that it may be overridden by later and better evidence. But, if so, this will not be a matter of the past event's having been prevented; and so we have no use for a present-past sense of the past tense. All we shall ever have concerning the past is *evidence*; there are no present tendencies toward something's having happened.

To deny the reality of the past is to think that, once it has passed away, all that remains of it are the traces it has left behind. It is not as if it had never been, since there are those traces: but of that of which no trace survives, it is as if it had never been. One who thinks this and also holds that the future is not real believes either that no statement in the future tense is true, or that, if it is, it is true by virtue of the clues that foreshadow the event predicted or surmised. He thinks, therefore, that all there *is* is the ever-changing present. The justificationist believes that a statement can be true only in virtue of evidence for it, whether anyone is in actual possession of that evidence or not; but, under the proposed modification, he allows that, when the statement is about the past, that evidence may itself lie in the past. He is therefore no longer disposed to deny the reality of the past. He is supported by the greater fidelity of this view to our linguistic dispositions: our disinclination to admit a present-past tense, and our view of present traces as no more than evidence for a past event.

How is it, then, concerning the reality of the future? A notable exponent of the view that the past is real, but the future is not, was C. D. Broad. He maintained in his book *Scientific Thought*[2] that the passage of time is a continual growth in the sum total of reality, which consists of what there is and what there has been. When an interval of time has passed, reality has increased by just that amount; as a moment that was future becomes present, reality acquires a new surface that it did not have before. A ground sometimes offered for this model of time is that we cannot in principle know the future. The reason offered for this is that, however high the probability of some belief of ours about the future, that belief cannot be *knowledge* because the possibility will always remain open that, at the time to which it relates, things may not happen in accordance with it. But, if a statement about the future cannot in principle be known to be true or false, then it cannot *be* true or false in advance of the time it is about: there can therefore be nothing before that time in virtue of which it is true or false. The presupposition of this reasoning is that every true statement can in principle be known to be true, or at least that, of any class of true statements, at least some member can in principle be known to be true.

This argument exemplifies a mistake frequently made in philosophical discussions about time. A philosopher wishes to demonstrate an asymmetry between past and future by instancing some thesis which holds good of the past but whose dual does not hold good of the future. His example is framed in terms of some temporally asymmetric expression, that is, one that is explained or understood by reference to a particular temporal direction. Plainly, such an argument begs the question: the alleged asymmetry was imported by the vocabulary in which the example was framed. In the present instance, temporally asymmetrical constraints have been placed on knowledge. It is not held that nothing can be knowledge if it is possible that there be contrary evidence stronger than the grounds on which it rests, but only that nothing can be knowledge if it is possible that contrary evidence stronger than the grounds on which it rests should *subsequently* come to light. We are interested in a concept of knowledge characterized in this temporally asymmetrical manner because we are interested in our own cognitive progress, as we are interested in our own moral improvement. I do not wish to deny the relevance of such a concept of knowledge to our lives; but it is far from obvious why this concept of knowledge should be linked to the concept of truth. We could appeal to a temporally symmetrical constraint on knowledge to support the conclusion that we can never *know* anything about the past: the possibility will always remain open that, at the relevant time, there was counterevidence stronger than the evidence we now have.

---

[2] New York: Harcourt Brace, 1923.

In her inaugural lecture at Cambridge, the late Elizabeth Anscombe[3] said, "if [my] physical movements are physically predetermined by processes which I do not control, then my freedom is perfectly illusory" (*ibid.*, p. 146). Since the scholastic period, it has been thought a problem to reconcile freedom of action with divine foreknowledge, indeed with there being an *actual* future course of events. The past is fixed, but the future is fluid: were it not, we could not affect the future, just as we cannot affect the past, and it would be senseless for us to ponder what we shall do, since it would *already* hold good that we should do one thing or another. This illusion derives from the idea that the present truth must compel the future action. The efficacy is in the reverse direction, however: a proposition about what I am going to do is true in virtue of my later action. Christ's prediction that Peter would deny him did not compel Peter to do so: it was true because he did. Indeed, if I *knew* what I was going to do, and knew that I knew, I could not deliberate over whether or not to do it: but I could consider whether or not I should give my will to doing it. Even if I knew that I was going to do something, and at that time give my will to doing it, I could consider whether or not I should *now* give my will to it. But the mere supposition that there is now a truth of the matter about what I am going to do is no reason for me not to deliberate, if I do not know what that truth is. Suppose I am playing a card game for very high stakes. After the hands have been dealt, I am told that I once played a hand, which I have now forgotten, in which all the players held just the same cards as now, but not told whether I won or lost. This experiment has been tried on many people; almost all play their hands as they did before. Knowing that it is almost certain how I am going to play does not inhibit me from thinking out my strategy, which I have as much reason to do as at any other time.

The same holds good for the claim that we cannot affect the past.

---

[3] *Collected Philosophical Papers*, Volume II (Minneapolis: Minnesota UP, 1981).

If I know what happened in the past, it would be senseless for me to try to prevent it from having happened, and even more to act so as to bring it about that it happened; and we are indeed seldom in this position with respect to what will happen in the future. But if I do *not* know what happened in the past, there is nothing senseless in my doing something to bring it about, or make it more probable, that some particular thing happened, if I think I know of some means of doing so. If someone prays that something should have happened, he is acting neither nonsensically nor blasphemously: he is not asking God to make something have happened which did not happen, but only that God should have made it happen.

There is a standard argument to show that, since a given past event has either happened or not happened, it must be either redundant or fruitless to do anything to bring it about that it has happened. If the event has happened, it will be redundant; if it has not happened, it will be fruitless. It usually goes unnoticed that this is the precise analogue to an argument for fatalism, to show it either redundant or fruitless to do anything to bring it about that something should subsequently happen, an argument that almost everyone would reject as sophistry. The fatalist argument starts from the premise that the event either will or will not happen. We want to reply that it may be that it will happen precisely because of what we do now to bring it about: not knowing whether it will happen or not, we have no reason to refrain from doing what we believe makes it more probable. Likewise, if we knew of something we could do that would make it more probable that some event previously took place, but did not know whether it had taken place or not, we should have no reason to refrain from doing that. To retort that the past event has *already* either happened or not happened, whereas the future event has not *yet* happened or failed to happen is obviously to assume what it was intended to prove.

It may be replied that, concerning what may be affected by our actions, we cannot know what will happen independently of our intentions, whereas we can know, independently of our intentions, that some event has taken place in the past. Precisely so. If we had a strong belief in some means of bringing it about that something had previously taken place, then we should no longer feel able to be sure whether an event of that kind had taken place, independently of our intention to employ, or not to employ, that means. Causal efficacy does appear, in our experience, to run only in the earlier-to-later direction; this is a fact about the world, not a product of our conventional use of verbs such as 'cause' and 'bring about'.

If statements about the future acquire a truth value only at the time to which they relate, truth values are to be relativized, not only to the time when a statement is made, but also to the time of evaluation. If some event—say your scratching your ear—is happening now, the link between the truth values of statements made at different times demands that the statement, made a year ago, that you would be scratching your ear in exactly a year's time was *then* true, not merely that it would be true in a year's time. One who denies the reality of the future must therefore hold that it may be correct, at a given time, to assign a truth value to a statement in the future tense made previously but relating to that time. That earlier statement lacked a truth value when it was made: now it has acquired one. What provides any space for such relativization to time of evaluation? It may be replied that the denial of the reality of the past must envisage statements about the past as being *deprived* of truth values they formerly possessed, whereas the denial of the reality of the future requires statements about the future to *acquire* truth values that they formerly lacked. We may demand that a truth value, once acquired, should never be lost; we cannot demand that one can never be attributed if it has not always been possessed.

Why this difference between acquisition and deprivation? Yet again the question has been begged. The only difference between a proposition's being deprived of a truth value and its acquiring one consists in the temporal order of its possessing and lacking a truth value. To insist that the difference affects the principle that the truth value of a statement is independent of when that truth value is ascribed to it is to assert no more than that, in one case, we are concerned with the future and, in the other, with the past: it is not to give any reason why that should be relevant.

The standard arguments for the unreality of the future—or however that thesis should be formulated to allow for the fact that 'the future' is not an absolute term—are to be rejected. For a justificationist, direct evidence for a statement that an observable event will take place at some future date must be an observation made at the time to which it relates (or when such an observation is possible); such a statement will be true if such an observation is or could be made. The analogy between the past and future tenses is complete.

It thus seems that time is just one of the dimensions along which reality varies, as it does along any of the three spatial dimensions; our awareness is one of the respects in which it varies. There is therefore no distinction between past and future in reality; there is no *now* in reality. There is indeed a distinction between one temporal direction and the other; the earlier-to-later direction is that of causal efficacy. Hence we assign to the past those events capable of having a causal influence upon events near us, so that we can receive information from them and of them, but have no means of affecting them; and we assign to the future those events which we can affect, but from which we can receive no information. Time is thus not merely one dimension out of four; but what occupies it in each direction is equally real.

Does not this four-dimensional conception exemplify the error of *spatializing* time? John McTaggart[4] argued that to deny that being past, present, or future is an objective feature of reality is to delete what is essential to time—change. On the four-dimensional conception, there is no real change: there is only "Cambridge change," the analogue of saying, 'The landscape changes as you travel east'. The landscape does not change: it is simply different here and east of here. The four-dimensional conception is grounded on the idea of our consciousness moving along the temporal dimension through a static four-dimensional reality. Obviously, the conception is incoherent. Time has not been reduced to one of four dimensions; it remains as the measure of the movement of our consciousness along one dimension.

A proponent of the four-dimensional conception may deny this. We are, he says, irregular four-dimensional tubes (or hypertubes), with the peculiarity that consciousness attaches only to our temporal cross-sections. Nothing changes: it is just that our different temporal cross-sections are aware of different things. But consider a description of other hypertubes, whose axes lie along a spatial dimension. To us these would appear long, very short-lived objects; if we learned that a different consciousness attached to each segment of one of the tubes, we should regard them as strings of distinct creatures. But if we were told that a different consciousness attached to each cross-section of such a tube in a direction orthogonal to its axis, and changed continuously in the direction of the axis, we could make nothing of this at all. The description of ourselves as hypertubes only appears to make sense when interpreted as 4D-speak for a single consciousness moving along the time dimension. The four-dimensional conception remains incoherent.

The reality of past and future does not, in fact, entail this picture of reality. Objects—galaxies, stars, planets, plants, and animals—

---

[4] "The Unreality of Time," in his *Philosophical Studies*, S. V. Keeling, ed. (New York: Longmans, 1934), pp. 110–31.

change as they move through space; when the object is a sentient creature, its experiences and observations of its surroundings change; when it is a rational creature, its thoughts and beliefs, its conception of its surroundings, change also. We are wrong if, on the strength of the reality of past and future, we replace in our conception of the world enduring and changing three-dimensional objects by static four-dimensional ones. What picture, then, should we have of the universe as a whole as it is in itself rather than from any point of view within it? The question is nonsensical; there is no such thing as a picture from no point of view. Physical reality consists in the objects in the world, the changes they undergo, and the paths they trace in space-time.

We are strongly disposed to think of simultaneity as an absolute relation; indeed, I earlier spoke of a child's grasping that he can ask, 'What is happening there now?', as an essential step in his coming to understand statements about other places. Yet there is no compelling rational ground for the absoluteness of simultaneity. Imagine that there were a permanent dense fog separating the Old World from the New, and that no regularity was detectable in the time taken to travel from one to the other: sometimes only a few weeks would have passed in the other hemisphere after a traveler's last visit there, sometimes many years. It would then be senseless to ask what had been happening in one half of the Earth at the very time that something happened in the other, as we might feel an urge to do; yet each half would have its history.

To defend the determinateness of reality, and the principle of bivalence, realists sometimes appeal to God's omniscience, or at least to the conception of an omniscient being. Such an appeal is certainly fruitless unless there really is an omniscient being. It falls little short of fatuous to argue that, even if there is not an omniscient being, certain truths must hold good because they would have held good if there had been an omniscient being. At first sight, there actually existing an omniscient being can have no bearing on what truths there are, other than the truth of the proposition that such a being exists and others entailed by it. Whatever truths there are, God, being omniscient, will know them; if there are no truths of a certain kind, he will not know any such. If there are gaps in reality—questions that have no true answer—then God will not know any anwers to such questions: he knows all that there is to know, but not more than there is to know. But this is to overlook the self-reflexive character of knowledge. If God knows a truth, then he also knows that he knows it; if he does not know something, then he also knows that he does not know it, and hence that it is not true. To object to the 'also', on the ground that a truth's holding good and God's knowing that it does are the very same thing, would be irrelevant. What is relevant is that God must know *whether* he knows that such-and-such a thing holds good.

From this it appears that, as the realist maintained, omniscience implies bivalence. God knows every true proposition; and if God does not know, of some proposition, that it is true, then he knows that it is *not* true. So it seems that classical logic is the logic of God's knowledge. For us to use classical logic is to yield to the temptation to behave as if we were gods, who know every truth and repudiate every falsehood.

It is a defective appearance, however. The deduction of bivalence from omniscience is erroneous: this is not bivalence as the realist hoped to establish it. A proposition is true if there are grounds to be had for it, whether or not we are in possession of them. A proposition is false if its negation is true; on the justificationist understanding of negation, the negation is true if there is an obstacle to there ever being grounds for the proposition. This is why there can be gaps in reality: the absence of such an obstacle does not entail that there *are* any grounds for the truth of the proposition. Now, if there is such a gap, there will be a proposition of which God does not know either its truth or its falsity. He knows that the proposition is not among the things he knows. It follows that the proposition fails to be true, but not, of course, that its negation, as we understand negation, is true. The proposition's failing to be true covers two cases, its being false and there being a truth-value gap; but this is not the sense that we can give to the expression 'not true'. We cannot recognize any specific gap in reality: we are entitled to declare a proposition not to be true only if we recognize an obstacle to its being so, and hence if we recognize it to be false. God can know where such gaps occur, namely, when he knows neither the truth nor the falsity of some proposition; he has available to him a negation which is not available to us. It might, therefore, perhaps be said that the logic of God's knowledge is a three-valued one, to whose application we cannot attain. Justificationist semantics requires us to use intuitionistic logic: it cannot be ruled out that we may find grounds for any proposition that we have not recognized as being false. We may sometimes be able to recognize that a proposition cannot be false, without yet being able to pronounce it true; but we cannot ever determine that such a proposition is not true.

MICHAEL DUMMETT

University of Oxford

MEMO TO AUTHORS

The Journal of Philosophy is published monthly, in issues that (ordinarily) contain only 56 pages of copy; this limits the length of papers the editors can normally publish to no more than 7,500 words, or 18 *Journal* pages. A more precise way of determining length is by characters (including spaces): 2,850 characters per *Journal* page, or 51,300 characters per *Journal* article of 18 pages.

The editors request that three copies of all manuscripts submitted be prepared according to the following guidelines:

(1) All material, including the quotations and footnotes, must be typed, double-spaced, and one-sided, with ample margins left and right, top and bottom.
(2) No bibliographies will be published; bibliographical material should be put into footnotes. No appendices will be published either.
(3) Footnotes should be complete, according to *Journal* style (see any issue). Journal citations should include: name of journal in full, volume number, issue number, month or season, year, and pages. Book references should include: publisher, U.S. city of publication, and date. Footnotes should be gathered at the end of the article.
(4) Quotations should be checked by the author; capitalization, italics, spelling, and the like should match the original.
(5) If there are diagrams, they should be provided camera-ready by the author. If there are special symbols, they should be widely spaced and clearly marked. Standard logic symbols are available, and letters of the Greek and script alphabet.
(6) No boldface or underlining; please limit use of italics to technical terms (when they are introduced) and to variables.
(7) No contractions.

The editors are willing to read and evaluate a manuscript that does not meet these guidelines, provided the author agrees to meet them before submitting a final manuscript. The editors may ask for revision of excessively long manuscripts, and may require reformatting of accepted manuscripts. They also request an electronic version with the final hard copy.

The Journal of Philosophy does not publish papers that are primarily historical or expository; the editors urge that such manuscripts be submitted elsewhere. While we do not publish manuscripts that mainly comment on articles published elsewhere, we do have a "Comments and Criticism" section in which short responses to *Journal* articles may be published.

Manuscripts may be sent to: John Smylie, Managing Editor, *The Journal of Philosophy*, 709 Philosophy Hall, 1150 Amsterdam Avenue, Mail Code 4972, New York, NY 10027.

The Center for Philosophy of Religion at the University of Notre Dame announces five fellowships for the 2003-2004 academic year two Postdoctoral Fellowships (stipend of $35,000 teaching one course each semester, otherwise $30,000), a Distinguished Scholar Fellowship (stipend of $50,000 teaching one course each semester; otherwise $45,000), an Extraordinary Fellowship ($40,000) to a scholar who might not qualify for the fellowships above, but who would benefit from a year at the Center—for example, foreign scholars or those outside the field of philosophy (e g , theologians who want to enrich their understanding of philosophy of religion and theistic philosophy, humanists from other fields who could benefit by a deeper grasp of the sort of philosophy practiced at the Center, and scholars interested in a theistic approach to psychology, sociology, economics and other subjects), and a Visiting Graduate Fellowship ($14,000) to a philosophy graduate student at another institution who is working on a dissertation in philosophy of religion or Christian philosophy and who would profit from spending a year at the Center

Applicants should see that the Center receives each of the following

(1) A complete curriculum vitae, which should include a list of publications and names of three people you will ask to send letters of recommendation on your behalf.
(2) Three letters of recommendation
(3) A statement of no more than three pages (double-spaced) describing the Fellowship for which you are applying and the project on which you would like to work while at the Center
(4) One published or unpublished paper.

All material should be sent to Thomas P Flint, Director, Center for Philosophy of Religion, 418 Malloy Hall, Notre Dame, IN 46556

The deadline for Extraordinary Fellowship and Distinguished Scholar Fellowship applications for 2003-2004 is January 15, 2003; for all other Fellowships, April 1, 2003.

E-mail: cprelig.1@nd edu Web site: http://www.nd edu/~cprelig/

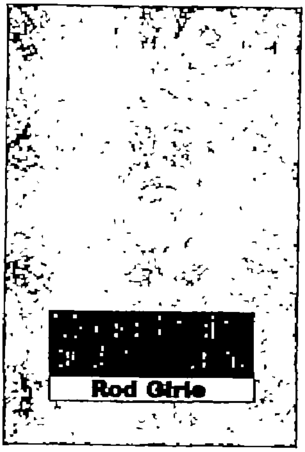
THE JOURNAL OF PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org.*

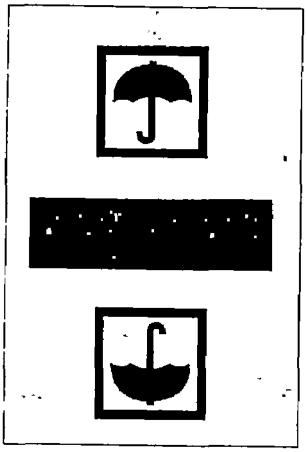
PURPOSE & HISTORY
CONTACTS AND ADDRESSES
BOARD OF TRUSTEES
EDITORIAL BOARD
MEMOS TO AUTHORS
SUBSCRIPTIONS & ADVERTISING
RECENT & FORTHCOMING ISSUES
ELECTRONIC ISSUES
INDEX 1904–PRESENT
RIGHTS AND PERMISSIONS
LECTURES & CONFERENCES
PHILOSOPHY LINKS

McGill-Queen's

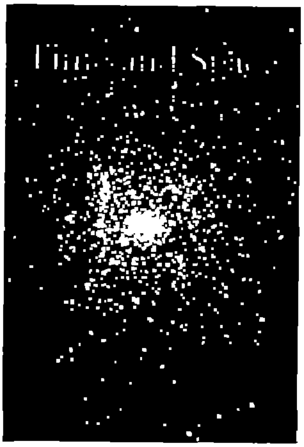
NECESSARY
Adoptions



0773521496 $22.95



0773524460 $19.95

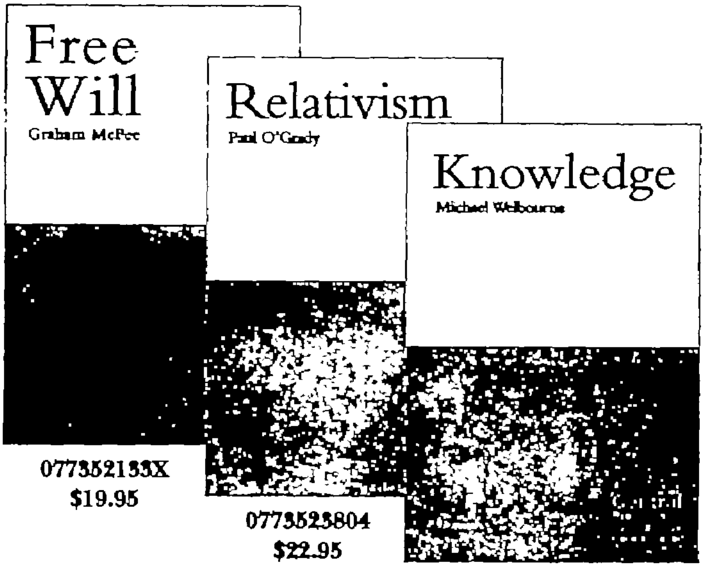


0773523065 $22.95

0773523073 $19.95

mqup.ca

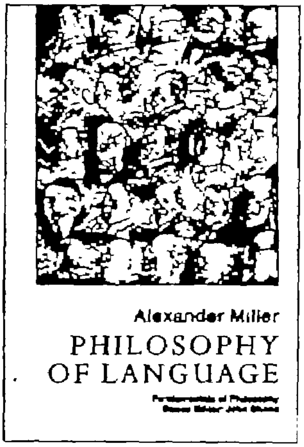
University Press



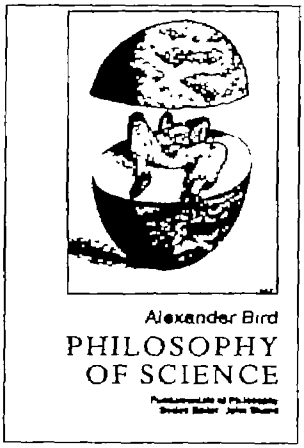
077352133X
$19.95

0773523804
$22.95

0773523057 $22.95



077351709X $19.95



0773517731 $22.95

Distribution services through
Cornell University Press

mqup.ca

Philosophy

NEW FROM CHICAGO

ENCOUNTERS AND REFLECTIONS

Conversations with Seth Benardete

Edited by Ronna Burger

By turns wickedly funny and profoundly illuminating, *Encounters and Reflections* presents a captivating and unconventional portrait of the life and works of Seth Benardete. With vignettes about fellow students and colleagues who were to become major intellectual figures, this book is the closest thing we will have to an autobiography of one of the 20th century's leading intellectuals.

CLOTH $30.00

EXILING THE POETS

The Production of Censorship in Plato's *Republic*

Ramona A. Naddaff

"With great originality and verve, this book explores the ways in which philosophy's spokesperson inherits the mantle of the poets even as he banishes them from the ideal city. Naddaff produces passages that are simply brilliant in their new perspective on a very old problem as well as in their insight and lyricism."—Shadi Bartsch, author of *Ideology in Cold Blood* and *Actors in the Audience*

CLOTH $27.50

THE GRACE AND THE SEVERITY OF THE IDEAL

John Dewey and the Transcendent

Victor Kestenbaum

"The status of Dewey scholarship has been increased immeasurably with this book. Informed by wide-ranging research on the sundry interpretations of Dewey's thought, Kestenbaum rescues Dewey's pragmatism from the stereotypical confines of instrumentalism and naturalism by locating a vibrant dynamic of transcendence in his works on ethics, aesthetics, and religion."—Calvin O. Schrag, The George Ade Distinguished Professor of Philosophy Emeritus, Purdue University

PAPER $20.00

HOW SCIENTIFIC PRACTICES MATTER

Reclaiming Philosophical Naturalism

Joseph Rouse

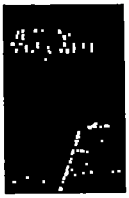
"A bold and very satisfying work. Rouse tackles head-on some of the central problems animating current philosophical discussion, in Continental circles as well as in Anglo-American 'analytic' philosophy, and in science studies as well as in philosophy proper."
—Thomas Nickles, University of Nevada

CLOTH $49.00

THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu

ROUTLEDGE

NEW IN PHILOSOPHY



ANNOUNCING A NEW SERIES

ROUTLEDGE AMERICAN PHILOSOPHY SERIES (RAPS)

Series Editor: John J. Stuhr, Penn State University

The *Routledge American Philosophy Series* presents original philosophical work advancing traditions of American thought so as to address new intellectual issues and global problems, as well as historical and theoretical studies of pragmatism, classical American philosophy, and related developments in American thought and culture.

A COMMUNITY OF INDIVIDUALS
John Lachs
Drawing on the spirit of classical American Philosophy, this book is an original, highly engaging, and important philosophical exploration of pressing personal and social problems including the complex relationships between individuality and community.
232 pp • $21.95

PRAGMATISM, POSTMODERNISM AND THE FUTURE OF PHILOSOPHY
John J. Stuhr
In this energetic and far-reaching new book, Stuhr draws persuasively on the resources of the pragmatist tradition of James and Dewey, and critically engages the work of Continental philosophers like Adorno, Foucault, and Deleuze, to explore fundamental questions of how we might think and live differently in the future.
224 pp • $22.95

FORGIVENESS AND REVENGE
Trudy Govier
"Splendid. Cogently and persuasively argued, it makes out the case for forgiveness and reconciliation on sound philosophical and secular grounds. I am very impressed."
—Archbishop Desmond Tutu
224 pp • $22.95

READING MCDOWELL
On Mind and World
Edited by Nicholas Smith
"McDowell's path-breaking Mind and World *was a brilliantly original, provocative, and rather cryptic book. Reading McDowell makes it much easier to grasp his intentions and strategies. It contains stimulating and fruitful exchanges between McDowell and some of the major figures of contemporary philosophy."*
—Richard Rorty, Stanford University
328 pp • $26.95

THE IMPORTANCE OF BEING UNDERSTOOD
Folk Psychology as Ethics
Adam Morton
Morton examines key concepts in moral psychology including motive, belief and simulation, arguing that folk psychology plays a vital role in building co-operation between individuals.
240 pp • $29.95

THE LIFE OF THE MIND
An Essay on Phenomenological Externalism
Gregory McCulloch
In a spirited and rigorous attack on most of the orthodox positions in contemporary philosophy of mind, McCulloch connects three of the orthodoxy's central themes—externalism, phenomenology and the relation between science and commonsense psychology in a defense of a thoroughly anti-Cartesian conception of mental life.
176 pp • $25.95

1.800.634.7064
www.routledge-ny.com

Routledge
Taylor & Francis Group

Surface and Depth

Dialectics of Criticism and Culture

RICHARD SHUSTERMAN

"In the 1960s, artists became concerned with closing the gap between art and life. Philosophers responded by insisting that such a gap is essential if art is to have any meaning. In this masterful and informed philosophical critique, Richard Shusterman shows how his pragmatist aesthetics can dissolve the dilemma and realign the relationships among philosophy, art, and life."—Arthur C. Danto, Columbia University. $49.95 cloth, $19.95 paper

The Art of Life

JOHN KEKES

Reflecting on lives of integrity and honor, Kekes formulates what we can learn from them and what we can do to adapt the ideals they represent to our personal circumstances. Avoiding both the abstractness that characterizes much moral thought and the relativism that recognizes no rational or moral limits, Kekes shows how serious philosophical thinking can be readable and helpful to those who struggle with the perennial problems of human existence. $29.95

Reinventing Pragmatism

American Philosophy at the End of the Twentieth Century

JOSEPH MARGOLIS

"In this often brilliant treatment of the first- and second-wave pragmatists, Joseph Margolis displays a knack for weaving together a complex history and an important contemporary philosophical debate. Margolis's analyses are bold, original, and intellectually exciting."
—Armen T. Marsoobian, Southern Connecticut State University, Editor in Chief, *Metaphilosophy*. $39.95

The Art of Humane Education

DONALD PHILLIP VERENE

"This is a vital and radically important book. Donald Phillip Verene engages education in the humanities at the most fundamental level and does so in language that is both clear and profound. *The Art of Humane Education* addresses major issues—tradition, cultural change, the idea of a canon, the differences between the humanities and the sciences, practical problems in teaching—and does so in a way that achieves earned wisdom."—James Engell, Harvard University. $18.95

An Essay on Divine Authority

MARK C. MURPHY

In the first book wholly concerned with divine authority, Mark C. Murphy explores the extent of God's rule over created rational beings. The author challenges the view—widely supported by theists and nontheists alike—that if God exists, then humans must be bound by an obligation of obedience to this being. He demonstrates that this view, the "authority thesis," cannot be sustained by any of the arguments routinely advanced on its behalf. After exposing the inadequacies of the various arguments for the authority thesis, he develops his own solution to the problem of whether, and to what extent, God is authoritative. Cornell Studies in the Philosophy of Religion. $39.95

The Suspicion of Virtue

Women Philosophers in Neoclassical France

JOHN J. CONLEY, S.J.

The salon was of particular importance in mid- to late-seventeenth-century France, enabling aristocratic women to develop a philosophical culture that simultaneously reflected and opposed the dominant male philosophy. In *The Suspicion of Virtue,* John J. Conley, S. J., explores the moral philosophies developed by five women authors of that milieu: Madame de Sablé, Madame Deshoulières, Madame de La Sablière, Mlle de la Vallière, and Madame de Maintenon. 5 illustrations. $39.95

A Politics of Impossible Difference

The Later Work of Luce Irigaray

PENELOPE DEUTSCHER

The influential philosopher and theorist Luce Irigaray has been faulted for giving more importance to sexual difference than to race and multiculturalism. Penelope Deutscher's eagerly awaited book, the first to focus on the scholar's controversial later works, addresses this charge. Through a learned critique of these lesser-known writings, the book examines Irigaray's claim that the politics of feminism and multiculturalism are intrinsically linked. The volume also serves as a clear and comprehensive introduction to her entire corpus. $39.95 cloth, $17.95 paper

NEW IN PAPER

"I'm Not a Racist, But . . ."

The Moral Quandary of Race

LAWRENCE BLUM

North American Society for Social Philosophy "Social Philosophy Book of the Year" Award

"This book combines clarity and compassion to an extraordinary degree, while surveying in an accessible way a vast range of thinking about race. It will be enormously useful to teachers . . . It will advance the debate among those who think about race; but above all, its wide diffusion will benefit all of us as citizens of a racialized society."
—Kwame Anthony Appiah, Harvard University. $16.95

Making Sense of Taste

Food and Philosophy

CAROLYN KORSMEYER

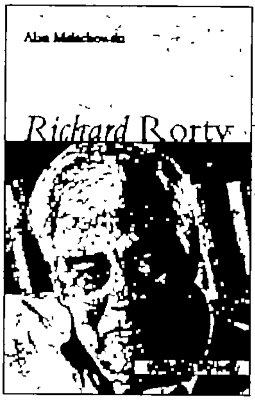
"Of the five senses, two—sight and hearing—were higher and lent themselves to aesthetic perception, while the remaining three—touch, taste and smell—were lower and non-aesthetic senses. Korsmeyer, in this sensitive and judicious book, explores and exposes the errors misinforming this conventional ranking."—*British Journal of Aesthetics.* 18 illustrations. $19.95

At bookstores,
or call (607) 277-2211

Cornell University Press

www.cornellpress.cornell.edu

New from Princeton



Richard Rorty
Alan Malachowski

Richard Rorty is one of the world's most influential living thinkers. Here Alan Malachowski provides an authoritative overview of Rorty's considerable body of work and a general assessment of his impact both within philosophy and in the humanities more broadly.

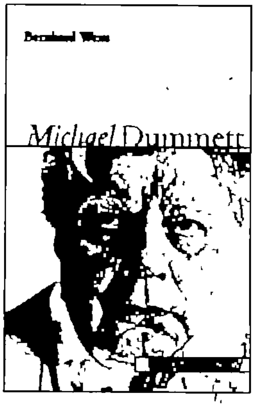
"This is an engaging, timely, and useful book. It offers the student and general reader something that is badly needed a wide-ranging, accessible and sympathetic introduction to Rorty's thought."—Michael Williams, Johns Hopkins University

Philosophy Now Series
Paper $17.95 ISBN 0-691-05708-7 Cloth $55.00 ISBN 0-691-05707-9
Available from Princeton only in the Americas

Michael Dummett
Bernhard Weiss

No one seriously interested in the philosophy of language can afford to ignore the work of Michael Dummett. This book offers a systematic presentation of the main elements of Dummett's pivotal contribution to contemporary philosophy. Its overarching theme is his discussion of realism: Bernhard Weiss explores the philosopher's characterization of realism, his attack on realism, and his invention and exploration of the anti-realist position.

This book is a sympathetic and accessible study, aiming not only to expose but to engage both with Dummett's philosophical thought and with his philosophical character.



Philosophy Now Series
Paper $18.95 ISBN 0-691-11330-0 Cloth $55.00 ISBN 0-691-11329-7
Available from Princeton only in the Americas

Princeton University Press

THE JOURNAL OF
PHILOSOPHY

VOLUME C, NUMBER 2
FEBRUARY 2003

19/5/03

page
55 *The Role of the Priority Rule in Science* Michael Strevens
80 *Quantum Life: Interaction, Entanglement, and Separation* Eric Winsberg and Arthur Fine

BOOK REVIEWS
98 Robert Stalnaker: *Context and Content* William W. Taschek

109 NOTES AND NEWS

UNIVERSITY LIBRARY

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721

You may also visit our website at www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER. Periodical postage paid at New York, NY, and other mailing offices.

POSTMASTER. Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 2, FEBRUARY 2003

THE ROLE OF THE PRIORITY RULE IN SCIENCE

Science aims to provide goods that all can share, but it does so in an atmosphere that is as competitive as it is cooperative. Consider what happens when several research programs all pursue the same scientific goal: there is a competition that has the form of a winner-takes-all race, that is, a race in which there are no second prizes.

Remarkably enough, this is true in two quite distinct ways: first, with respect to the benefit that a successful research program confers on society, and second, with respect to the personal rewards—fame, prizes, authority, and so on—that accrue to the members of a successful research program. Let me say more about each.

First, benefit to society: it is a consequence of the very nature of scientific information that, while the first of several competing research programs to achieve the programs' common goal can expect to confer perhaps a great deal of benefit on society, the runners-up will confer very little or nothing. The reason is, of course, that additional discoveries of the same fact or procedure are pointless. Thus, whereas a typical economic goal, such as providing a swordfish steak or a back rub, can be realized over and over again to society's benefit—we all have use for more than one meal or massage—the goal of a scientific research program needs to be realized just once for society to benefit maximally. Competing research programs are therefore racing to be the provider of a benefit in the knowledge that the first to achieve their common goal will, in reaching it, nullify the worth of the efforts of all the other competitors. Call this the *benefits race*.

Second, personal benefit: possibly the most distinctive feature of the social organization of science is the *priority rule*, the system of

0022-362X/03/0001/55–79

rewards which accords all credit, and so all the personal benefits that go along with credit, to the first research program to discover a particular fact or procedure, and none to other programs pursuing the same goal. As a consequence of the priority system, workers in competing research programs are involved in a winner-takes-all race for personal rewards. Call this the *rewards race*.

I will argue that the strong parallel between the payoff structure of the benefits race and that of the rewards race, that is, between the scheme by which science benefits society and the scheme by which science rewards its practitioners, is not accidental: the fact that scientific rewards are distributed according to the winner-takes-all priority system, rather than some other scheme, is explained by the winner-confers-all manner in which science benefits society. The structure of the benefits race, then, explains society's implementation of the rewards race.

The form of the explanation is as follows. A scientific reward scheme such as the priority rule acts as a system of incentives, encouraging researchers to devote their time and energy to some research programs in preference to others. Different reward schemes, then, may result in different allocations of resources among competing research programs. Society has an interest in adopting a reward scheme that promotes an allocation with a relatively high expected payoff. I will show that the priority system promotes an especially efficient allocation of resources in "winner-confers-all" situations, that is, in situations where almost all benefit is extracted from a goal the very first time it is reached.

I. THE RULE OF PRIORITY

I.1. Aspects of the priority rule. The reward system to which I refer as the priority rule can be divided into two parts. First, rewards to scientists are allocated solely on the basis of actual achievement, rather than, for example, on the basis of effort or talent invested. Second, no discovery of a fact or a procedure but the first counts as an actual achievement. The second part gives the rule its name, but the first part is, I think, equally worthy of note.

A qualification: it is not strictly true to say that the priority rule governs the distribution of all rewards in modern science, since most scientists receive some kind of salary, regardless of their achievement. Although it is possible to finesse the rule by holding that the basic salary of, say, an academic scientist is compensation for teaching and administration, rather than for scientific endeavor, I will for simplicity's sake put aside the question of monetary reward altogether, concentrating on rewards that take the form of reputation, a sizable office, the rapt attention of graduate students, and the like—in short, the rewards that take the form of *prestige*. Prestige is, it is generally agreed, allocated more or less in accordance with the priority rule.

It was the sociologist of science Robert Merton[1] who established the priority rule as a characteristic of the social organization of science worthy of serious study. Merton's discussion, still an essential source for any research on priority, seeks to establish a number of claims about the priority rule of which I single out three as especially important.

First, Merton argues that the concern for priority and for actual achievement is not so much a consequence of the psychological makeup of scientists as the reflection of a powerful norm governing the scientific endeavor. Scientists, then, are not merely psychologically disposed to allocate prestige in accordance with the priority rule; they feel obliged to do so—they regard a distribution of prestige in accordance with the rule as right and proper, and a distribution that violates the rule as wrong or unjust. So, for example, Merton holds that disputes over priority are to be explained principally as a consequence of the moral concern of the scientific community that the rules be enforced, not as a consequence of a clash of individual egos.

Second, Merton documents a concern with the priority rule going back to the beginnings of modern science, that is, to the early seventeenth century. Galileo and Newton are cited as early and eager parties to priority disputes that arose from a concern with the rule; according to Merton, the rule has held sway ever since. The priority system, then, seems always and everywhere to have determined the allocation of credit in Western science.

The third Mertonian observation that I wish to consider here concerns the extreme literalness with which the priority rule is enforced: if the same fact is discovered twice, Merton notes, the first discovery garners all the rewards no matter how slender the margin by which it edges out the second. To this effect, Merton quotes the permanent secretary of the French Academy of Sciences and distinguished physicist François Arago, writing in the first half of the nineteenth century:

> Questions as to priority may depend on weeks, on days, on hours, on minutes.

---

[1] "Priorities in Scientific Discovery," *American Sociological Review*, xxii (1957). 635–59, reprinted in *The Sociology of Science* (Chicago: University Press, 1973), pp. 286–324 Page references are to the latter version.

This attitude, Merton writes, "only [puts into] words what many others have expressed in behavior" (*ibid.*, p. 322). He cites a dispute between Galileo and Simon Mayr over the observation of Jupiter's moons in which the relevant time interval was a mere ten days (*ibid.*, p. 287).

Call this phenomenon—the fact that no finish is so close that the priority rule ceases to apply—the *Arago effect.* Merton regards the Arago effect as a pathology of the social organization of science, a "dysfunctional extreme far beyond the limits of utility" (*ibid.*, p. 322). His view is shared by others writing on the priority rule, either explicitly, or implicitly in the sense that their explanations of the rule do indeed accord no advantage to a concern with days, hours, minutes. A feature of my own explanation of the priority rule is that it does find a constructive role for the Arago effect to play: worrying about a matter of ten days between discoveries can exert a positive influence on the way in which science is conducted.

I.2. Alternatives to the priority rule. There are a number of salient alternatives to the priority system; three will provide sufficient contrast for my purposes here.

(1) Scientists might all, on an egalitarian reward system, receive equal compensation—that is, for the purposes of this paper, equal prestige.
(2) Scientists might be compensated in proportion to the talent and effort they invest in their research. Harder working, cleverer scientists would receive more prestige.
(3) Scientists might be compensated in proportion to their actual achievements, in accordance with the priority rule, but with the amendment that being the second to make a discovery, or having progressed considerably towards making a discovery, count as actual achievements. Runners-up in a scientific race would, in other words, receive prestige for finishing well.

The question with which I am concerned, then, may be put thus: why does the scientific community disburse prestige in accordance with the priority rule rather than one of the above reward schemes, or some alternative scheme?

The aspect of the question that is perhaps most puzzling concerns the contrast between system (2), which rewards participants according to talent and effort invested, and the priority rule and system (3), which reward participants according to actual achievement. I say this because the factor that makes the difference between talent and effort, on the one hand, and actual achievement, on the other, would appear to be simple luck. Two equally brilliant scientists invest equal time in pursuing some goal; one happens to choose the right method and so succeeds, the other does not. What reason can there be to concentrate all rewards on the scientist who makes the lucky choice? Any truly satisfying explanation of the priority rule must answer this question.

With an eye to this question and others, I propose the following three-way classification of reward systems. First, there are egalitarian systems, which distribute rewards more or less equally, such as scheme (1). Second, there are what I will call resource-based schemes, which distribute rewards in proportion to the resources invested by the rewardee. Scheme (2) is a resource-based scheme. Third and finally, there are achievement-based schemes, which distribute rewards in proportion to the achievements of the rewardees. Both scheme (3) and the priority rule are achievement based. This classification is not intended to be exhaustive; in section II, I will describe a system that seems to fit somewhere between resource- and achievement-based schemes.

The main questions I seek to answer in this study can now be posed as follows:

(1) What explains the fact that science's reward system is achievement based rather than resource based?

(2) Given that science has an achievement-based system, why do second and subsequent attainings of a goal not count, for the purpose of determining rewards, as achievements?

I.3. Explanations of the priority rule. Merton and others after him have explained the primacy of the priority rule in a number of ways. Let me sketch four such explanations.

(1) The priority system gives scientists an incentive to undertake research, both by rewarding them directly for their research and by creating the perception, through the rewards, that discovery has its own intrinsic value.[2]

(2) It is much easier to measure achievement than to measure the talent and effort invested in a research program. Furthermore, it is difficult to assess the progress made by the runners-up once a scientific race has been won, since it is easy for runners-up to steal the now-published results of the winners. Thus, the priority system is far more practical than reward systems such as schemes (2) and (3) above.[3]

(3) The priority rule gives scientists an incentive to publish their re-

---

[2] See Merton.

[3] Partha Dasgupta and Paul A. David, "Toward a New Economics of Science," *Research Policy*, XXIII (1994): 487–521.

search as soon as possible, making the benefits promptly available to society (*ibid.*).

(4) It may simply seem fairest to reward scientists in exact proportion to the benefits they confer on society, perhaps especially so to the members of a market-driven society.

All four explanations have their merits. They also have drawbacks, but this is not the place to investigate the drawbacks. None of the four can, by itself, explain all aspects of the priority rule considered above. This advantage I claim for my explanation alone, to which I now turn.

II. RESOURCE ALLOCATION: THE ADDITIVE CASE

II.1. The problem of resource allocation. Every scientific research program might usefully be better funded, better staffed, better equipped. When new resources, pecuniary and human, come available for science, there will always be a real question as to how best to distribute them among research programs, both those that exist and those that ought to exist. This is the problem of resource allocation.

There is a wide and a narrow sense of the problem. In the wide sense, all research programs are competing for money and personnel. But especially fraught—and this creates the problem in its narrow sense—is the competition between research programs aiming to realize the same goal. This paper considers both problems, first the problem of distributing resources among research programs with different goals, in this section, and then the problem of distributing resources among programs competing to reach the same goal, in section III.

It is the contrast between the solutions to the two kinds of problems that will provide the basis for my explanation of the priority system. Any reward scheme functions as a system of incentives, directing researchers towards some programs rather than others. A reward scheme, then, has a considerable effect on the allocation of labor among research programs. I will show that the priority system is apt to achieve an especially efficient allocation of labor in a winner-confers-all benefits race, and so an especially efficient allocation of labor among research programs competing to achieve the same goal. (Readers seeking a better sense, ahead of time, of how the argument goes might turn to section III.2.)

II.2. A simple model of the resource allocation problem. The question of resource allocation in science was raised quite some time ago by Charles Sanders Peirce;[4] it has recently been brought once again to philosophers' attention by the multifaceted investigations of Philip Kitcher.[5] My simple mathematical model for studying resource allocation is borrowed, along with the idea that there might be a connection between resource allocation and reward schemes, from Kitcher.

---

[4] "Note on the Theory of the Economy of Research," in Arthur W. Burks, ed., *Collected Papers of Charles Sanders Peirce*, Volume 7 (Cambridge: Harvard, 1958), pp 76–83.

The starting points for the model are the following assumptions, which are of course quite idealized:

(1) Every research program has a single goal. There are only two possible outcomes of the program's endeavors: total success, if it realizes the goal, or total failure, if it does not.
(2) Different research programs have different intrinsic potentials.
(3) A program's chance of success—that is, the probability that it will achieve its goal—depends on two things, its intrinsic potential and the resources invested in the program.

Given these assumptions, a research program's probability of success can be written as a function of the resources invested in it. I call this function the program's *success function*, and I write it $s(n)$, where n quantifies the resources invested.

One research program has more intrinsic potential than another if, given any fixed level of investment, the one has a higher chance of success than the other. For simplicity's sake, I assume that there is a complete ordering of programs in terms of their intrinsic potential, or equivalently, that the success functions for two programs do not intersect unless they entirely coincide. This is also true of Kitcher's model.

The problem of resource allocation is the problem of distributing limited resources among different programs so as to maximize the return to society. As Peirce and Kitcher have noted, it is not in general true that the optimal allocation devotes all resources to the program with the greatest intrinsic potential; the problem, then, is not trivial.

In order to pose the problem more precisely, it is necessary to write down a function expressing the value to society of any given distribution of resources. To this end, suppose that for any research program, there is some fixed utility to society of that program's success achieved in isolation. Writing the utility v, the expected return from investing n resources in a program is the product of the utility of success and the probability of success, that is, $vs(n)$.

Now suppose that there are two research programs, with utilities v_1 and v_2, and success functions $s_1(\cdot)$ and $s_2(\cdot)$. What is the expected return to society from investing n_1 resources in the first and n_2 resources in the second? In the simplest case, it is the sum of the individual expected returns, namely,

---

[5] "The Division of Cognitive Labor," this JOURNAL, LXXXVII, 1 (January 1990): 5–21; and *The Advancement of Science* (New York: Oxford, 1993).

$$v_1 s_1(n_1) + v_2 s_2(n_2)$$

I call this the *additive case*. All other cases are *nonadditive*.

Let me give an example of an additive case and an example of a nonadditive case. Expected returns from two programs will be additive if the goals of those programs are, in an intuitive sense, independent. If one program seeks a cure for cancer and another the Higgs boson, for example, then benefits are independent and returns are additive.

Returns are nonadditive in many circumstances, but the most important for my purposes is the case in which (a) both programs have the same goal, and (b) second and subsequent realizations of that goal are of no value to society, creating a winner-confers-all benefits race. In the winner-confers-all case, the expected return is the utility v of the goal in question multiplied by the probability that at least one program succeeds, that is,

$$v(s_1(n_1) + s_2(n_2) - s_{12}(n_1, n_2))$$

where the third term $s_{12}(n_1, n_2)$ is the probability that both programs achieve their goals, given the allocation of n_1 resources to the first and n_2 resources to the second. It is the existence of this third term that makes the expected returns nonadditive.

Because I am interested in winner-confers-all races, it is the nonadditive case that ultimately concerns me. However, as I said above, what concerns me more specifically is the contrast between the optimal allocations of resources in the additive and the nonadditive cases. For this reason, I investigate the optimal allocation for the additive case in the remainder of this section.

II.3. Optimality in the additive case. You have a limited number N of resources—say, worker hours—to distribute between two research programs. Your goal is to maximize the total expected return. The programs' benefits are additive. Assuming that additional worker hours always improve a program's chances of success, you will distribute all N worker hours available, n to one program and $N - n$ to the other. Your return, then, is

$$v_1 s_1(n) + v_2 s_2(N - n)$$

The problem is to find the value of n that maximizes this return.

For simplicity's sake, assume that v_1 and v_2 are equal. (Alternatively, assume that the utilities are built into the success functions, so that

$s(n)$ represents not the probability of success but the expected return, given an investment of n worker hours.) Then what must be maximized is the expected number of successes:

$$s_1(n) + s_2(N - n)$$

This is the kind of constrained maximization problem that is familiar from economics. In solving the problem, I make one very important assumption, that the success functions yield *decreasing marginal returns*, that is, that each additional worker hour invested in a program increases the probability of success a little bit less than the last (though there is always some increase, as assumed above).

The notion of a marginal return on investment will be an important one here, so let me present a definition and some notation to smooth the discussion. The marginal return from investing one worker hour in a research program, given that n worker hours have already been invested, is the increase in the probability of success brought by the investment of that additional hour, that is,

$$s(n + 1) - s(n)$$

I write this as $m(n)$. To say that marginal returns are decreasing is to say that the function $m(n)$ is strictly decreasing, that is, that the graph of $m(n)$ slopes downwards from left to right.

The assumption of decreasing marginal returns is a familiar one; I will not try to justify it here (though I note that it may have its exceptions, as suggested by Kitcher[6]). From an economic point of view, the assumption has a pessimistic feel: a constant rate of progress requires ever-increasing investment. But from a mathematical point of view, decreasing marginal returns are a wonderful thing, because the assumption of decreasing marginal returns makes possible extremely general claims about constrained maximization problems such as the resource allocation problem.

In the additive case, the assumption of decreasing marginal returns implies that the function to be maximized will have a single maximum point corresponding to the value of n at which the marginal returns are equal, that is, at which

$$m_1(n) = m_2(N - n)^7$$

It is not guaranteed that this value of n is realistic, that is, nonnegative, but provided that N is large enough—and I assume throughout that it is—the guarantee can be made. The result generalizes in a number of ways, some of which are described at the end of section II.4.

---

[6] "The Division of Cognitive Labor," p. 13.

[7] Because the marginal return function is an approximation to the derivative of the success function.

If the additive version of the resource-allocation problem is thought of as a static problem in central planning, then its solution is straightforward. Simply allocate the N worker hours available to you, the central planner, so that the marginal returns from each of the programs, on the investment of a further worker hour, would be about equal.

The same solution can be used for dynamic central planning, that is, central planning where a constant stream of worker hours is coming available, and must be allocated as they arrive so as to provide as high a possible return on investment over time.

To see this, note that returns over time will be maximized if, for any N, over the interval where exactly N worker hours, and no more, have become available, they are distributed in such a way as to maximize returns. Provided that the central planner always distributes worker hours so as to keep the marginal return functions as equal as possible, the static result guarantees that this condition will be satisfied for any N, and so the dynamic problem is solved.

This treatment of the dynamic problem makes two rather significant assumptions: first, that a program's success function, and in particular, its intrinsic potential, does not change over time, and second, that the central planner knows at all times the true form of the success function. These assumptions are obviously oversimplifications; the issue is further discussed at the end of section II.4.

II.4. A reward scheme for the additive case. In the last section, I imagined that resource allocation is to be administered by a single, all-powerful central planner who has only the interests of science and society at heart. But in fact, with respect to worker hours at least, allocation in science is driven to a great extent by certain decisions of individual scientists, namely, their decisions as to what projects to pursue. These decisions are, I presume, strongly influenced by scientists' expectations of rewards, in particular, the complex of rewards that I am calling prestige; it is at this point that reward schemes at last make their entrance.

Let me simplify the issues by making the following two assumptions: first, that the distribution of worker hours among research programs is entirely determined by the choices of the workers themselves, and second, that the workers' choices are entirely determined by the prevailing reward system, so that they choose to allocate their time to the project that will bring them the greatest expected return. I am imagining, then, that every researcher periodically—at the end of a subproject, say—reconsiders his commitment to his current program, staying on only if the reward for devoting his next, say, year to that program is at least as great as the reward for investing the year in some other program. If he does not stay, he moves, of course, to the program that will reward him best for his year's service.

It is important to note that the programs themselves do not distribute the relevant rewards. The programs may pay salaries, but they have no power to bestow prestige. That is a power of the scientific community as a whole, to be exercised in accordance with the prevailing reward system, an aspect of what Merton calls the "normative structure of science."

From the perspective of the problem of resource allocation, an ideal reward system is a scheme of rewards that encourages scientists to allocate their labor in accordance with the precepts laid down in the last section, that is, so as to keep the marginal returns from every program as close as possible.

This can be done, obviously enough, by encouraging scientists always to invest their time in the program that, at the moment of their decision, offers the greatest marginal return. Given decreasing marginal returns, this will eliminate, as quickly as possible, the differences between the marginal returns from different programs. An optimal reward scheme, then, is one that, at any given time, makes the program with the highest marginal return the most attractive.

The simplest such scheme awards prestige in proportion to marginal return. That is, it bestows prestige in proportion to the amount by which scientists increase their program's probability of success, or more generally, the amount by which they increase their program's expected utility to society.[8]

---

[8] Let me situate this reward scheme in a more general framework. Define as a program's *reward function* the function $r(n)$ that yields the reward given to a person who invests a worker hour in the program at the point at which n worker hours have already been invested. For example, if you join an program in which i worker hours have been invested, you receive the equivalent of $r(i)$ utility points. Suppose that there are a number of different programs, each with its own reward function, and suppose that each worker always invests his next hour in such a way as to maximize his reward. Then, provided that the reward functions are *strictly decreasing*—that is, provided that the more worker hours have been invested in a program, the lower the reward to a worker who invests one more—workers will distribute their investment among the programs so that, at any particular time, the disparity between the rewards that would be obtained by investing one more hour in any of the programs is as small as possible, that is, so that the reward functions are, at any time, approximately equal. To ensure that marginal returns are kept approximately equal, then, set the reward function for a program proportional to its marginal return function. More generally, to ensure that any other set of functions, such as the adjusted marginal return functions to be described in section III.1, are kept approximately equal, use

Call this scheme the *Marge* reward scheme. *Marge* is not an achievement-based scheme, because it rewards contributions to the probability of success rather than successes themselves. But it is not a purely resource-based system either, or, at any rate, scientists' rewards are not determined solely by the talent and effort that they invest.[9]

In addition to talent and effort, what is rewarded is timing. Under a regime of decreasing marginal returns, an hour invested earlier in a program's life brings a greater marginal return than an hour invested later. The reward for the earlier hour will be commensurately greater than the reward for the later hour. Thus, two scientists of equal talent and industry, working within the same program, will not be rewarded equally. The scientist whose career unfolds earlier in the program's history will make a greater contribution to the program's probability of success, and so will be rewarded with greater prestige than his psychological double.

For expository purposes, I nevertheless classify *Marge* as a resource-based scheme, on the grounds that rewards are based entirely on a scientist's input, regardless of the output. Think of the value of the resources provided by a scientist as determined not only by the intrinsic properties of those resources—by talent and effort—but also by timing.

In summary, then, there is a reward system, the *Marge* scheme, which will ensure an optimal allocation of labor among any set of research programs, provided that the programs have goals with additive values and yield decreasing marginal returns on investment. This is a result of considerably greater generality and precision than has previously been stated in the literature on resource allocation in science.[10]

But the result has its limitations, due to the assumptions and idealizations I have made along the way. Some of these can be relaxed without compromising the result. I have already noted that the result

---

those functions as the reward functions.

[9] My model does not represent talent explicitly, but it is easy to build talent into the representation by counting an hour of work from a talented scientist as worth more than one "standard" hour.

[10] Kitcher, for example, examines only one particular kind of success function, and suggests a reward scheme—an equal division of some fixed prize among the participants in a successful research program—that will normally result in a somewhat suboptimal allocation of labor. Kitcher's main contribution is not a particular mathematical result, but rather the insight that different scientific reward systems will result in different solutions to the resource allocation problem, a sine qua non of my own work.

is true for cases in which there are more than two competing research programs, and that the effect of talent is easily incorporated into the model (see footnote 9). Two other respects in which the assumptions can be loosened without lessening the optimality of the *Marge* scheme are as follows.

First, it is possible to assign to each program a number of goals, rather than a single goal, each with its own probability of realization. Provided that these goals are additive, and that scientists are rewarded in proportion to the sum of their contributions to the realization probabilities of each of their program's goals, *Marge* will find the optimal allocation of labor.

Second, it is possible to incorporate the effect of positive feedback among different research programs, such that an investment in one program increases the probability not only of that program's success but also of other programs' successes. Provided that a scientist's reward recognizes her contribution not only to her own program, but to all other programs as well, *Marge* will steer scientists so as to achieve the best possible allocation of labor.

There are other assumptions that cannot be relaxed without compromising *Marge*'s optimality. If, for example, the success function of a program can change in unforeseeable ways, optimality may be out of reach. But this is a state of affairs that equally hinders the wisest central planner. Because investment of time is irrevocable—hours spent in research cannot be retrieved and reassigned—it is always possible that a change in a success function will create a situation in which previous resources turn out to have been invested so badly that it is impossible, with the resources in hand, to attain the optimal allocation. The problem, then, is not a consequence of any peculiar fragility of the reward scheme that I have suggested, but is rather a universal obstacle to achieving the best possible outcome when information is limited.

Indeed, the *Marge* strategy will do as well as a central planner could under the same circumstances: it will move to accommodate unanticipated changes in the success functions as fast and efficiently as possible. Although *Marge* may not achieve the optimal allocation of resources in such a case, then, the allocation is as good as can be expected given the information available. The same comment applies to any case in which knowledge of the success function is limited.

These are perhaps not very exact claims, but precision is, in any case, about to be abandoned. The investigation of winner-confers-all races in the next section will require a move from absolute statements of the form *this reward scheme is optimal* to comparative, qualitative statements of the form *of these two reward schemes, this one is better.* It is the comparative, qualitative statements that will form the basis for the explanation of the priority rule.

III. RESOURCE ALLOCATION: THE WINNER-CONFERS-ALL CASE

III.1. The winner-confers-all benefits race. In a classic scientific winner-confers-all benefits race, two research programs compete with one another to make the same discovery. When one program succeeds, society receives the full benefit of the discovery; whether or not the other program later also succeeds makes no difference to society's welfare.

As stated in section II.2, society extracts the greatest return from a winner-confers-all race when the allocation of resources among programs maximizes the probability that at least one program succeeds. If it is assumed, for expository simplicity, that the successes of two competing programs are independent (nothing important will turn on this assumption), then what must be maximized is

$$s_1(n_1) + s_2(n_2) - s_1(n_1)s_2(n_2)$$

The optimal distribution of N worker hours, then, is that in which the number n of worker hours allocated to the first research program maximizes

$$s_1(n) + s_2(N - n) - s_1(n)s_2(N - n)$$

Define the marginal return function $m(\cdot)$ for a program as before, so that $m(n)$ is the increase in the probability of success due to an additional worker hour being invested on top of n hours already invested. Thus, $m(n) = s(n + 1) - s(n)$, as in the additive case. Then the optimal value of n will be the value for which

$$m_1(n)(1 - s_2(N - n)) = m_2(N - n)(1 - s_1(n))$$

I call the functions on either side of the equals sign the *adjusted marginal return functions.*

What I want to emphasize about this solution to the winner-confers-all allocation problem is not its exact form, which after all depends on the perhaps dubious assumption that the success of one program is independent of that of the other, but rather its qualitative properties. Specifically, what is important is that, compared with the optimal distribution for the additive case, the optimal distribution for the winner-confers-all case allocates more workers to the program with the higher intrinsic potential.[11] This observation is generally true, both for cases where more than two research programs compete to achieve the same goal, and for cases where the successes of different programs are not independent.[12] (To avert any confusion: the optimal distributions for both additive and winner-confers-all scenarios allocate more resources to the program with the higher intrinsic potential. The distribution for the winner-confers-all scenario, however, favors the higher-potential program more heavily than does the distribution for the additive scenario.)

It follows that the optimal distribution in a winner-confers-all race is different from the optimal distribution in an additive scenario. Because the optimal distributions differ, the *Marge* reward scheme will not be optimal in the winner-confers-all case. A new reward scheme is needed.

III.2. Three reward schemes. The optimal distribution of resources in a winner-confers-all benefits race allocates relatively more resources than the *Marge* scheme to a program with higher potential. I now show how the *Marge* scheme can be transformed so as to create the necessary additional bias in favor of higher-potential programs. The effect of the transformation is to turn *Marge* into science's priority system.

The transformation has two steps. First, *Marge*'s resource-based scheme is transformed into an achievement-based scheme on which scientists are rewarded only if their program achieves its goal. Priority plays no role in this scheme: if two rival programs both achieve their common goal, scientists in both programs are rewarded. I call this scheme (or rather, the specific version of the scheme that I describe in section III.3) the *Goal* reward scheme.

The second step of the transformation turns *Goal* into a scheme in which there is a priority race, so that, if two rival programs both achieve their goal, only the first to do so is rewarded. I call the resulting scheme the *Priority* reward scheme.

The two new schemes will retain *Marge*'s principle for dividing rewards among a program's workers: scientists are rewarded in proportion to their contribution to their program's probability of success. What changes are the rules for deciding when programs are to be rewarded at all. The reward schemes become progressively less generous: *Marge* rewards all programs, *Goal* rewards only programs that achieve their goals, and *Priority* rewards only programs that are the first to achieve their goals.

---

[11] To see this, note first that the greater the relative value of a program's marginal return function, the more resources are allocated to that program, and second, that compared to the marginal return functions, the adjusted marginal return functions will be relatively lower for programs with relatively lower intrinsic potentials, due to the adjusting factor's negative dependence on the success probability of the competing program. (By the adjusting factor, I mean the $(1 - s_2(N - n))$ factor.)

[12] There are exceptions to this latter claim in extreme cases, but the extreme cases are not realistic.

In section III.3, I will show that *Goal* favors higher-potential programs relatively more than does *Marge*, that is, that it attracts relatively more workers to higher-potential programs. In section III.4, I will show that *Priority* favors higher-potential programs relatively more than does *Goal*. Moving in the direction of a priority system, then, affects the allocation of resources so as to better suit a winner-confers-all benefits race. Whether *Goal* or *Priority* achieves a more efficient allocation in a winner-confers-all race is not, however, a question that I will be able to answer here, for reasons to be explained in section IV.1.

III.3. From a resource to an achievement basis. The *Marge* scheme is transformed into the *Goal* scheme by rewarding only those programs that achieve their goal. I will, in this section, make one further assumption about the workings of *Goal*, as a consequence of which it will turn out that a scientist's expected reward from any investment is the same on the *Goal* scheme as on the *Marge* scheme (down to a constant of proportionality). Yet *Goal* achieves a different allocation of resources from *Marge*, I will go on to argue, because of human *risk aversion*.

The further assumption I make about *Goal* is that there is a fixed amount of prestige bestowed on each program that is rewarded, to be divided among the program's workers according to their contributions. A scientist's share of this prize, then, is equal to her probabilistic contribution to her program's success as a proportion of the program's total probability of success. This means that, although workers in programs with relatively lower intrinsic potentials have a relatively lower chance of receiving any reward at all, when they are rewarded, they will receive a relatively larger share of the prize, since lower-potential programs have lower success probabilities.

These two factors exactly cancel out, so that on *Goal*, the expected reward for a scientist's investing in a particular program is proportional to the probabilistic contribution that the investment makes, in absolute terms, to the program's success (with the constant of proportionality the same for all programs). In other words, the expected value of investing in a program is proportional to the reward that the scientist would receive on the *Marge* scheme for the same

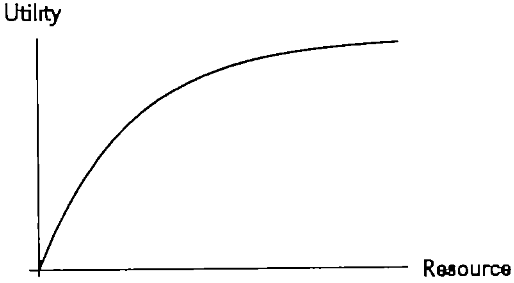


Figure 1: The relationship between the utility and the quantity of a resource, for an agent who is risk averse with respect to the resource

investment. As a consequence, if scientists were to make investment decisions solely on the basis of the expected reward from their investments, *Goal* would result in exactly the same preference ordering among programs in all situations as the resource-based *Marge* scheme. Either scheme, then, would result in the same distribution of worker hours among research programs.

To undermine this unwelcome conclusion, I appeal to risk aversion. I use the notion of risk aversion in its technical sense. An agent is risk averse with respect to some kind of resource, such as money or social influence, if that resource has *decreasing marginal utility* for the agent, meaning that, when the agent has a certain quantity of the resource, a small amount more is not worth as much to him as it would be if he had less of the resource. The qualitative relation between the quantity of the resource and the utility is shown in figure 1.

If a person is risk averse with respect to, say, money, he will think and behave in certain characteristic ways. For example, he will consider a loss of a given value to be more undesirable than a gain of the same value is desirable. He will reject some fair gambles. And, most important for what follows, when two equally priced lottery tickets have the same expected monetary value, but one has a higher probability of winning (which means it must have a lower payoff), a risk-averse decision maker will prefer the higher probability ticket. For example, given a choice between playing a game in which, if a tossed coin lands heads, you win $1000 (there is no gain or loss if the coin lands tails), and a game in which, if the single red ball is drawn from an urn containing 100 balls, you win $50,000, if you are risk averse, you will choose the coin toss game. To put it qualitatively, you prefer a large probability of winning a small prize to a small probability of winning a large prize.

It is this preference that allows an achievement-based reward scheme to increase the relative appeal of higher-potential research programs. Consider a scientist choosing between two rival research programs that have different potentials but that are currently offering equal expected rewards. On the *Marge* scheme, the scientist is sure to be rewarded whichever program she chooses, so she will be indifferent between the two. On *Goal*, the scientist is more likely to be rewarded by the higher-potential program than by the lower-potential program. A risk averse scientist will therefore choose the higher-potential program. For this reason, more scientists will join higher-potential programs on *Goal* than on *Marge*, as desired in a winner-confers-all benefits race.

III.4. Implementing a priority race. What effect will moving to a priority race have on an achievement-based scheme such as *Goal*? To answer this question, consider the relation between a scientist's expected reward on the *Goal* scheme and the same scientist's expected reward on *Priority*, a reward scheme identical to *Goal* except that it rewards only the first program to achieve any given goal.

Suppose, in particular, that a scientist is choosing between two research programs that would seem equally attractive on the *Goal* reward scheme. The first of these programs has the higher intrinsic potential, let us say, and so a higher probability of reaching its goal, but the second has few enough workers that the additional share of the reward a scientist in that program would receive, were the program to realize its goal, exactly makes up for the lower probability of being rewarded. (Assume that the effects of risk aversion are also taken into account.)

How would the scientist's preferences change if the *Goal* scheme were to be replaced, at this moment, by the *Priority* scheme? The change will not affect the size of the reward the scientist is liable to receive. Its impact is confined to the probability of the scientist's being rewarded.

In a two-program race, *Goal* rewards a scientist in three circumstances:

(1) The scientist's program achieves its goal but the other program does not.
(2) Both programs achieve their common goal; the scientist's program achieves the goal first.
(3) Both programs achieve their common goal; the scientist's program achieves the goal second.

The *Priority* scheme rewards the scientist in situations (1) and (2) above, but not in situation (3). Because the situations are mutually exclusive, the probability of a reward on *Priority* is equal to the probability of a reward on *Goal* less the probability that situation (3) occurs.

Upon moving to *Priority*, then, the probability of a reward will decrease for both programs. The amount of the decrease is the probability of situation (3), which is $s_{12}w_2$ for the first program and $s_{12}w_1$ for the second program, where s_{12} is the probability that both programs achieve their goal (given the allocation of resources at the time of the decision), and w_1 and w_2 are the probabilities that the first and the second programs respectively will win the priority race, in the event that both achieve their goals.

It seems reasonable to suppose that, in almost all cases, the program with the higher probability of achieving its goal—in the example, the first program—is also more likely to win a priority race. Thus, w_1 is greater than w_2. It follows that, upon a move from *Goal* to *Priority*, the probability of the first program's being rewarded decreases by less than does the probability of the second program's being rewarded. Consequently, the expected reward from an investment in the first program will decrease by less than the expected reward from an investment in the second.[13] The scientist who is undecided under *Goal* will therefore choose the first program—the program with the higher potential—under *Priority*. (Risk aversion exacerbates this effect.) More generally, a move from *Goal* to *Priority* has the effect of increasing the bias in favor of higher-potential programs.

To summarize, in section III I have shown that

(1) To find the optimal allocation of resources in a winner-confers-all benefits race, a greater bias toward high-potential programs is needed than is created by *Marge* (section III.1), and

(2) *Goal* makes higher-potential programs relatively more attractive than *Marge* (section III.3), and *Priority* makes them more attractive still (section III.4).

IV. THE PRIORITY SYSTEM EXPLAINED

IV.1. From additive to winner-confers-all cases. I make the following conjecture: a resource-based reward scheme that distributes labor optimally in resource allocation problems that are additive, or nearly additive, can be adapted to work well in winner-confers-all benefits races by retaining the scheme's apparatus for calculating rewards while moving to an achievement basis for determining the recipients of the rewards. In a winner-confers-all race, then, figure rewards as before, but give the rewards only to programs with concrete achievements.

---

[13] The more so because the size of the second program's reward is larger.

This, I suggest, is the explanation, or at least a great part of the explanation, of the priority system in science. Our normal reward schemes, devised to handle problems of resource allocation that are close to additive, are modified for scientific resource allocation by adopting something similar to the previous section's *Priority* scheme.

The proposed explanation provides very satisfying answers to two questions about the priority system posed at the beginning of this paper concerning, first, the apparently excessive role of luck in the priority system's selection of reward recipients, and second, the status of the Arago effect.

The question about luck is raised by the observation that two equally talented, equally industrious scientists may receive different rewards under the priority system just because one of them, but not the other, is lucky enough to select a research program that achieves its goal (and does so before any rival). Previous explanations of the priority system have little to say about this facet of the system: they can only dismiss it as an unfortunate side effect of a system that works well in other respects. On my explanation, far from being a side effect, it is essential to the system's social function. By making rewards in part a matter of luck, the priority system harnesses humans' natural risk aversion so as to direct scientists to prefer more strongly than they would otherwise research programs with a high intrinsic potential for success, as required for an optimal solution to the problem of allocating resources in a winner-confers-all benefits race. The arbitrary aspect of reward distribution in the priority system, then, is the very motor of the system's being.

The Arago effect is the strict enforcement of the winner-takes-all rule, even when the interval of time between the winner's success and the runner-up's success is very short. It has the consequence that a scientist may accrue all the prestige due to *the* discoverer by making the discovery only a matter of weeks before his competitors.

When the priority rule discriminates scientific winners and losers in a race this close, Merton writes, "priority has lost all functional significance." The Arago effect is "a dysfunctional extreme far beyond the limits of utility" (*op. cit.*, p. 322). My analysis of the incentive structure imposed by the priority rule shows that this is not so. Discriminating between a winner and a loser in a very close scientific race, and rewarding only the winner, can benefit science and society for the same reasons that, more generally, the priority rule is a good thing.

This point can be appreciated by comparing a fully Aragoiste priority system, which rewards only the first research program to make a discovery, regardless of how close the finish, to a more relaxed system that declares a scientific race a draw if two rival programs achieve their goals within, say, six months of one another, and rewards the participants in both programs equally. I will not present the calculations here (the comparison can be made in the same way that *Goal* and *Priority* are compared in section III.4), but the Aragoiste scheme makes higher-potential research programs relatively more attractive, by a small amount, than does the relaxed scheme. Aragoism will therefore affect the distribution of labor among competing research programs, and so will, contrary to Merton's claim, have real functional significance.

This argument does not show, of course, that the Aragoiste system will achieve a more efficient allocation of labor than the relaxed system. Perhaps the Aragoiste system makes high-potential programs too attractive, so that the resulting distribution of labor overshoots the optimal assignment of resources to these programs. To show that the Aragoiste system is superior would involve some substantive quantitative assumptions about the degree and form of human risk aversion, and about the relation between a program's probability of its achieving its goal and its winning a priority race. (As mentioned in section III.2, the same caveat applies to the determination of the relative merits of the *Goal* and *Priority* schemes.) My conclusion, then, is not that the Arago effect is always beneficial, but that it has a kind of effect—an effect on the distribution of labor—that may be beneficial. It is a mistake, then, simply to class the Arago effect as a pathology.

IV.2. Society's grand reward system. I have tentatively explained the connection, asserted at the beginning of this article, between the winner-confers-all nature of the typical scientific benefits race and science's winner-takes-all reward system. The explanation is rather complex, and the connection is therefore much less straightforward than one might suppose. Has the symmetry in payoff structure between a winner-confers-all benefits race and a winner-takes-all rewards race turned out to be a mere coincidence?

A coincidence would be a pity, both because it is more exciting to explain a striking symmetry than to dismiss it, and because the symmetry seems to have some ethical resonance for us, as exhibited in the final explanation of the priority rule in section I.3, which makes the psychological assumption that, to us, it seems plainly fair to reward scientists precisely in proportion to the actual contribution they make to society.

In this section, I will recover the importance of the symmetry, by way of the following proposal:

(1) There exists in human affairs a grand reward scheme, which determines the distribution of certain kinds of rewards, including prestige, in winner-confers-all benefits races and in many other scenarios.
(2) The purpose of this grand reward scheme is to promote the optimal allocation of labor among different activities, including but not limited to scientific research programs, whose primary goal is to contribute to the collective, not the individual, good.
(3) The grand reward scheme sets a total reward for the participants in an activity in direct proportion to the actual contribution made by that activity to society This total reward is divided among the activity's participants in proportion to their contribution to the activity's probability of success, as on the *Marge* scheme.
(4) The priority scheme is nothing but the particular form that the grand reward scheme takes when applied to a winner-confers-all benefits race.

If all of this is true, then the symmetry between the benefits race and the rewards race in science is not a coincidence; rather, it is an instance of a more general symmetry that underlies the distribution of many social rewards, a symmetry mandated by the grand reward scheme.

Let me begin. To keep things simple, consider, as before, cases in which there are just two programs between which a fixed quantity of resources must be divided. Assume further that each program's goal is worth the same amount, in isolation, to society. In sections II and III, I distinguished two resource allocation problems, the additive and the winner-confers-all scenarios. In the additive scenario, if both programs succeed, society receives twice as much benefit as it would if only one program had succeeded. In the winner-confers-all scenario, if both programs succeed, society receives no more benefit than it would if only one program had succeeded.

These two cases can be seen as the end points of a spectrum of distributive problems. The cases intermediate between the additive and winner-confers-all scenarios are those where, if both programs succeed, society receives more than if only one program had succeeded, but less than twice as much. An example of such a case is a scenario in which two research programs are pursuing different treatments for the same disease. Suppose that 25% of all patients respond to both treatments, 25% to just the first treatment, and 25% to just the second treatment. Then; if just one program successfully develops the treatment it is pursuing, 50% of all patients can be cured, whereas if both programs succeed, 75% of all patients can be cured. The benefit from a second success is less than the benefit from the first success, but it is far from zero.

The resource allocation problems along the spectrum can be parametrized with a single number, which I call a problem's *additor*. The additor expresses the utility brought by a second success as a proportion of the utility brought by the first success. If v is the utility of a single success, then, and a is the additor, the utility of two successes is $(1 + a)v$. The additive case has an additor of one, the winner-confers-all case an additor of zero, and the medical example from the previous paragraph an additor of one-half.

I want to ask, first, what is the optimal distribution of labor for values of the additor partway between zero and one, and second, what kind of reward scheme might best find this distribution. Take the optimal distribution for the additive case as the baseline. In the winner-confers-all case, as I have shown, the optimal distribution favors high-potential programs considerably more than in the additive case. The optimal distributions for cases with intermediate additors fall, as one might expect, between these two extremes. They assign more resources to high-potential programs than does the optimal distribution for the additive case, but fewer resources than does the optimal distribution for the winner-confers-all case. The required bias towards high-potential programs, relative to the additive case, of course increases as the additor decreases.

How might a reward scheme implement this additor-sensitive bias? I showed in section III.4 that the difference in bias between *Goal* and *Priority* is due to the term $s_{12}w$, where w is, in a two-program competition, the probability that the rival program wins the priority race. Call this the *priority term*. Because the priority term is larger for lower-potential programs, and the *Priority* scheme in effect subtracts it from a program's probability of being rewarded, the effect of the priority term is to create an additional bias towards high-potential programs. Suppose that the size of the priority term was reduced by some constant proportion for all competing programs. Then the biasing effect would be smaller, because the reward probabilities after subtraction would be closer than they are in the case where the term is not reduced. One way to control the bias towards high-potential programs, then, is to implement a policy that has the effect of multiplying the priority term by $1 - a$, where a is the relevant additor. The higher the additor, the more the priority term is reduced, and so the smaller the bias towards high-potential programs.

To see how to achieve this effect, consider a program's expected total payout to its workers, which on *Goal* is just the size of the reward v multiplied by the probability of success. The effect of the priority term is to subtract $s_{12}wv$ from a program's expected payout. What is desired, then, is a policy that alters this effect so that it becomes a subtraction of $(1 - a)s_{12}wv$. The most straightforward such policy is to award the runner-up in a priority race a second prize of value av.

This means that the runner-up's payout is exactly proportional to the benefit society receives from the runner-up's success, with the constant of proportionality the same as for the winning program. In the medical scenario, for example, the first program to succeed creates a 50% cure rate; a second success increases the cure rate to 75%. Thus, the runner-up increases the cure rate by a factor of one-half, and so receives half the reward due to the winner, a proportion necessarily equal to the additor of 0.5.

I suggest that the system sketched in the last paragraph is, as I earlier intimated, close to the scheme used by society to determine the distribution of certain kinds of rewards, including prestige, for activities that benefit society as a whole. To put the proposal as simply as possible: society accords prestige and other rewards to both the winner and the runner-up of any benefits race in proportion to the social good resulting from their respective successes.[14]

In a winner-confers-all case, where the additor is zero, the runner-up contributes nothing and receives no reward. This is, of course, the priority rule. But now it can be seen that the priority rule is an extreme case of a much more general reward scheme that matches rewards to actual contributions to society. It is with reference to the grand reward scheme that the symmetry between the winner-confers-all aspect of scientific benefits races and the winner-takes-all aspect of the priority system is explained.

Perhaps even more interestingly, our sense that the symmetry is somehow fair, even just, is, I propose, rooted in the fact that the grand scheme that mandates the symmetry is, from the point of view of the resource allocation problem, a very good one. If this is true, our sense of fairness is in some sense designed in part to solve efficiently life's many resource allocation problems.

I conclude that it is quite possible that the priority system was not purpose-built for science. It is, rather, implicit in a much more general reward scheme designed to handle the allocation of labor among projects producing any social good, whether the good is knowledge based or not. This accounts for the appearance of a concern with priority at the very dawn of modern science. The system did not have to be constructed; it was already present.

---

[14] As I have described the grand reward scheme, it functions in the additive case identically to *Goal*. But it is *Marge*, not *Goal*, that finds the optimal allocation of resources in the additive case. There are several ways to modify the grand scheme so as deal with the problem, but this will have to be a topic for a sequel.

Why, then, does adherence to the priority rule seem to be a new and distinctive practice? Most social goods to which the grand reward scheme applies have additors substantially greater than zero, thus the corresponding reward schemes lack the winner-takes-all aspect of the priority system. What is new in science is the prevalence of the winner-confers-all benefits race; the extreme values of the additors for such races give an old system dependent on the additor for its form the appearance of something new.

MICHAEL STREVENS

Stanford University

QUANTUM LIFE: INTERACTION, ENTANGLEMENT, AND SEPARATION*

> Individual objects are not fixed but fluctuating, constantly responding to their surroundings, bundles of relationships, rather than settled points.[1]

When distinct quantum systems interact, they appear to lose their individuality. In almost all cases, once systems are interacting, they no longer have a proper (or "pure") quantum state at all. Moreover, the quantum state of the merged or "composite" system involves statistical correlations among the components that seem to go beyond anything obtainable from the systems that entered into the interaction, even when those systems become spatially distant from one another. Erwin Schrödinger[2] coined the term *entanglement* (*Verschränkung*) to describe this situation. Revising the old-fashioned marriage vow ("What God has joined together let no man put asunder"), entanglement seems to imply that not even God can separate quanta once merged. We ask here whether this is really so. Specifically, we ask whether entangled quantum systems are nevertheless separable, in the sense that the statistical correlations of the composite system can be obtained from statistics of the components.

We approach this question in the way that John Bell[3] and others have approached the issue of locality. We look for a clean mathematical demonstration to settle the issue one way or another. The mysteries and puzzles over quantum locality emerge from demonstrating that when the correlations of entangled systems violate the Bell inequalities, which is almost always, they cannot be explained by introducing suitable statistical states for the components. These are states that satisfy a precise condition of local action or "locality" and a condition

---

*We are indebted to Jon Jarrett, Don Howard, Nick Huggett, and other members of the Chicago Area Philosophy of Physics group, for sparking our interest in this topic, and for exchanging ideas and criticism. We thank Mathias Frisch, especially, for catching an error in an earlier 'proof'.

[1] T. M. Luhrmann, *Persuasions of the Witch's Craft: Ritual Magic in Contemporary England* (Cambridge: Harvard, 1989), p. 239.

[2] "Die gegenwärtige Situation in der Quantenmechanik," *Die Naturwissenschaften*, XXIII (1935): 844–49, and also "Discussion of Probability Relations between Separated Systems," *Proceedings of the Cambridge Philosophical Society*, XXXI (1935): 555–63.

[3] See, for example, *Speakable and Unspeakable in Quantum Mechanics* (New York: Cambridge, 1987).

0022-362X/03/0001/80–97

of statistical independence called "factorizability."[4] Here we ask the comparable question for separability. Can we introduce suitable statistical states for the components that allow the correlations of the composite system to be explained on the basis of statistics of the components? If the answer is "yes" then we regard the components of an entangled system as separable; otherwise not. Surprisingly, we show that even when the Bell inequalities are violated, which is where entanglement really shows up, the components are separable.[5] In carrying out this demonstration, we actually show something more. For the states we introduce are local (although not factorizable) and hence we can conclude that entanglement is compatible with both separability and locality. Thus during marriage, as well as before, quanta may lead a healthy life of their own.

This robust conception of quantum life has been contested. Indeed violations of the Bell inequalities have prompted a literature on the metaphysics of microscopic systems that flirts with some sort of metaphysical holism regarding spatially separated, entangled systems. The rationale for this behavior comes in two parts. The first part relies on a proof, due to Jon Jarrett,[6] that the experimentally observed breakdown of the Bell inequalities entails a breakdown of the conjunction of the constraints just mentioned. Jarrett called these two constraints *locality* and *completeness*. We prefer the terminology used above of locality and *factorizability*.[7] The first part of the rationale for metaphysical holism urges that only Jarrett's locality allows for "peaceful coexistence" between quantum mechanics and special relativity. Factorizability, it is suggested, must be jettisoned.

The second part of the rationale introduces *separability*, as above, which is regarded as the opposite, in essence, of holism. The argument here draws crucially on the lemma that separability implies factorizability. It would follow from this lemma, the Jarrett result, and the failure of the Bell inequalities that if we want to preserve locality, we must jettison separability, and embrace its opposite—metaphysical holism. Thus blame for the failure of factorizability is shifted to the failure of separability. In this way the correlations observed in entangled systems are seen as the sign of a deep (and perhaps mysterious) holism in nature.

---

[4] See, for example, N. David Mermin, "Quantum Mysteries for Anyone," this JOURNAL, LXXVIII, 7 (July 1981): 297–308; and Arthur Fine, "Antinomies of Entanglement: The Puzzling Case of the Tangled Statistics," this JOURNAL, LXXIX, 12 (December 1982): 733–47.

[5] In the sequel, we concentrate on correlation experiments whose outcomes violate the Bell inequalities. For the type of experiments considered here, satisfaction of those inequalities is sufficient for there being a local, factorizable model—see Fine, "Hidden Variables, Joint Probability, and the Bell Inequalities," *Physical Review Letters*, XLVIII (1982). 291–95. Every such model is separable. Hence, satisfaction of those inequalities implies separability.

[6] "On the Physical Significance of the Locality Conditions in the Bell Argument," *Noûs*, XVIII (1984) 569–89.

[7] Fine, "Correlations and Physical Locality," in P. Asquith and R. Giere, eds., *PSA 1980*, Volume 2 (E. Lansing, MI: Philosophy of Science Association, 1981), pp. 535–56

By establishing that merged quanta may lead a life of their own, we reject here the second part of the rationale, in particular, the lemma on which it relies. Properly conceived, there is no relation of logical entailment that flows from separability to factorizability. Our discussion here complements that of Tim Maudlin,[8] who also doubts that separability implies factorizability. Maudlin conjures up an imaginative separating scheme involving tachyons and he shows how superluminal signaling may accommodate the statistics for a special experimental configuration where factorizability fails. (Without the picture of superluminal signaling, one of us suggested a similar scheme with the idea of "randomness in harmony."[9]) In the tradition of the literature on Bell's theorem, we ask here directly whether certain kinds of structures can in principle support the quantum correlations, regardless of how they may be implemented. Thus we concentrate on investigating whether the statistical relations posited by schemes like Maudlin's can be represented in mathematical models of the experimental situation, without relying on intuitions associated with tachyons, or other physical mechanisms.

In showing how to construct local and separable models of correlation experiments whose outcomes violate the Bell inequalities, we show that holism, however deep or mysterious, is not a price that we have to pay for keeping locality. But what of nonlocal theories; for example, the Bohm theory? If separability is compatible with locality, is it also compatible with nonlocality? We do not address this issue directly since our concern here has been to assess the price of locality given the failure of the Bell inequalities. If our suggestion in section IV for understanding the quantum states is correct, however, then the Bohm theory, like quantum theory itself, is separable. Hence if

[8] *Quantum Nonlocality and Relativity* (Cambridge: Blackwell, 1994), see p. 98.

[9] See Fine, "Correlations and Physical Locality," and "Antinomies of Entanglement."

the Bohm theory is actually nonlocal, then violations of the Bell inequalities turn out to be compatible with nonlocal as well as with local models, each of which respects separability.[10]

I. BACKGROUND: THE EXPERIMENTAL SETUP

In the simple EPR-Bohm type experiments we consider here (2×2 experiments), a source emits pairs of particles (call them that for simplicity) in the singlet state ψ. After emission, the particles in each pair move off in opposite directions to spatially separate wings, the *A*- and *B*-wings. Each wing contains a detector assembly with two settings (1 and 2), corresponding to the measurement of spin orientation (in the plane perpendicular to the "path" of the particles). When the particles arrive in their respective wings, the detectors register either $+1$ or -1 (corresponding to spin up or spin down in the relevant direction). In the singlet state the probability is one-half at either wing for either $+1$ or for -1 to occur, regardless of the setting there. The probability for pairs of outcomes, the joint probability, is a simple function of the angles of orientation measured in each wing.

We want to model such a correlation experiment, by introducing a space Λ of basic "states" distributed according to some fixed probability measure on the space. (The states in Λ are often referred to as "hidden variables" but we agree with Bell that this is a misleading and silly terminology; so we avoid it here.) Each state $\lambda \in \Lambda$ determines a set of joint probabilities of the form:

$$p_\lambda{}^{AB}(x,y|i,j)$$

where A is the assembly in the A-wing, B is the other wing assembly, x and y are the measurement outcomes at A and B ($+1$ or a -1) and i and j are the settings (either 1 or 2) on A and B respectively. We require that the model recover the probabilities $P^{QM}(A_i = x, B_j = y|\psi)$ assigned by the quantum theory in the singlet state ψ as averages over the space Λ of basic states; that is, we require that

$$<p_\lambda{}^{AB}(x,y|i,j)>_\Lambda = P^{QM}(A_i = x, B_j = y|\psi)$$

Since the $p_\lambda{}^{AB}(x,y|i,j)$ completely determine the probabilities of all possible outcomes of all possible measurements, it is entirely natural to regard the entire set of these functions as the state of the composite system. This way of treating the state accords nicely with the quantum mechanical rule according to which the state function of a system is completely determined by the family of probability distributions for all observables of the system.

---

[10] See J.T Cushing et al., *Bohmian Mechanics and Quantum Theory An Appraisal* (Boston: Kluwer, 1996) for interpretive essays on the Bohm theory. The essay by M. Dickson, "Is the Bohm Theory Local?" pp. 321–30, challenges the nonlocality usually associated with Bohm

We can then introduce what one would naturally call the marginal probabilities at λ of individual outcomes at individual measurement stations.

$$p_\lambda^{A}(x|i,j) = \sum_{y} p_\lambda^{AB}(x,y|i,j) \text{ and } p_\lambda^{B}(y|i,j) = \sum_{x} p_\lambda^{AB}(x,y|i,j)$$

These marginals constitute the states in each wing.

Locality and factorizability. We are now in a position to define the two probabilistic constraints mentioned in the introduction. *Locality* is the constraint that at each state λ the marginal probabilities at A (respectively, B) do not in any way depend on the switch settings at B (respectively, A).[11] That is:

$$p_\lambda^{A}(x|i,j) = p_\lambda^{A}(x|i,1) = p_\lambda^{A}(x|i,2)$$

and

$$p_\lambda^{B}(y|i,j) = p_\lambda^{B}(y|1,j) = p_\lambda^{B}(y|2,j)$$

Assuming locality we can simplify the notation and write

$$p_\lambda^{A}(x|i,j) = p(A_i = x|\lambda)$$
$$p_\lambda^{B}(y|i,j) = p(B_j = y|\lambda)$$

for the local states.

The second constraint is *factorizability* (Jarrett's "completeness" and Abner Shimony's "outcome independence"). It is the constraint that at each λ the joint probabilities be the arithmetic product of their respective marginals. That is:

$$p_\lambda^{AB}(x,y|i,j) = p_\lambda^{A}(x|i,j) \cdot p_\lambda^{B}(y|i,j)$$

For experiments of the type considered here, the Bell inequalities constrain the difference D between the sum of any three joint probabilities (that is, the ensemble averages) and the fourth. The constraint is just that $0 \leq D \leq 1$. The experimental evidence is that, for certain settings of detectors in these experiments, the measured values for the joint probabilities violate the Bell inequality. Jarrett showed that if the inequalities were violated, then at most one of the above constraints, either locality or factorizability, could hold. At least one of them has to go. In these circumstances many think that the right strategy is to keep locality and to jettison factorizability. While that conclusion has important critics,[12] for the purposes of this article we will take it at face value. Our concern is how philosophers have come to interpret what they have perceived as the failure of factorizability. Below are some typical examples.

---

[11] Abner Shimony calls this "parameter independence"—see "Search for a Worldview Which Can Accommodate Our Knowledge of Microphysics," in J. Cushing and E. McMullin, eds., *Philosophical Consequences of Quantum Theory: Reflections on Bell's Theorem* (Notre Dame, University Press, 1989), pp. 25–37.

II. SEPARABILITY

> The fact that there exist quantum states of two-body systems which cannot be factorized into products of one-body quantum states means there is *objective entanglement* of the two bodies, and hence a kind of *holism*.[13]

> In terms of the Bell-type experimental setup, for example, we are barred from representing what we would ordinarily call the "two particles" in a given "pair" as two distinct entities; although detectable in spacelike measurement events, "they" form a single object, connected in some fundamental way that defies analysis in terms of distinct, separately existing parts.[14]

> [The source of the Bell result is] a kind of ontological holism or nonseparability..., in which spatio-temporally separated but previously interacting physical systems lack separate physical states and perhaps also physical identities.[15]

If we want to get clear on the relationship between factorizability and holism, we really need to understand holism's opposite, separability.

---

[12] See, for example, M. Jones and R. Clifton, "Against Experimental Metaphysics," in Peter A. French, et al., eds., *Midwest Studies in Philosophy*, Volume 18 (Notre Dame: University Press, 1993), pp. 295–316; Maudlin, p. 93ff; and J. Cushing, *Quantum Mechanics: Historical Contingency and the Copenhagen Hegemony* (Chicago: University Press, 1994), pp. 57–59.

[13] Shimony, p. 27

[14] Jarrett, "Bell's Theorem: A Guide to the Implications," in Cushing and McMullin, pp. 60–79, here p. 79.

[15] Don Howard, "Holism, Separability, and the Metaphysical Implications of the Bell Experiments," in Cushing and McMullin, pp. 224–53, here p. 225.

Don Howard's analysis (*ibid.*)[16] is an excellent place to start. In these works, Howard argues that the experimental violations of the Bell inequalities force us to abandon the fundamental metaphysical principle of separability and to embrace, instead, a kind of holism.

In trying to formulate a precise notion of separability, Howard draws inspiration from a passage of Albert Einstein's:

> It is characteristic of these physical things that they are conceived of as being arranged in a space-time continuum. Further, it appears to be essential for this arrangement of the things introduced in physics that, at a specific time, these things claim an existence independent of one another, insofar as these things "lie in different parts of space."[17]

Howard takes this "requirement" offered by Einstein as a fundamental metaphysical constraint on physical systems; the principle of separability.[18] He then offers a more precise formulation: the contents of any two spatiotemporally separated regions can be considered to constitute separate physical systems.[19]

Here *separate physical systems* means:

(1) Each system possesses its own *distinct* physical state.
(2) The joint state of the two systems is *wholly determined* by these separate states.

We find this articulation of the principle of separability to be very compelling.[20] It is, however, rather abstract. In particular, in order to apply this principle to a quantum system, one would need to articulate both what is meant by the *state* of the system, and what counts as a state being *wholly determined* by two other states. Howard does just this in a footnote, where he specifies how his principle is to be applied to quantum systems:

---

[16] See also his "Einstein on Locality and Separability," *Studies in History and Philosophy of Science*, xvi (1985): 171–201; and "Locality, Separability, and the Physical Implications of the Bell Experiments," in A. van der Merwe, F. Selleri, and G. Tarozzi, eds., *Bell's Theorem and the Foundations of Modern Physics* (Singapore: World Scientific, 1992), pp. 306–14.

[17] "Quantum Mechanics and Reality," *Dialectica*, II (1948): 320–24. This translation from the original German is by Howard, "Holism, Separability, and the Metaphysical Implications of the Bell Experiments," p. 223.

[18] Maudlin (*op. cit.*, p. 97) notes that Bell makes a similar assumption when he supposes that each region of space-time has an intrinsic state.

[19] Howard, "Holism, Separability, and the Metaphysical Implications of the Bell Experiments," p. 226.

[20] It does seem to us that there is a third condition that needs to be brought in, namely, that the states of separate parts are relatively independent of one another. (We say 'relatively' because there may be some conserved quantities shared by both systems whose values will be linked.) So minimally one would want that you cannot get *all* the information about one system part from the state of the other. Without something like this, separability would not necessarily run counter to holism (since the whole, then, might be fully reflected in both parts, in which case we would not need their 'sum'), which surely we want it to do. Most likely Howard's notion of 'distinct' partial states is intended to do this job—if worked out. None of the rest of our argument will hinge on this small quibble.

> How the joint state is determined by the separate states depends upon the details of a theory's mathematical formulation. At a minimum, the idea is that no information is contained in the joint state that is not already contained in the separate states, or alternatively, that no measurement result could be predicted on the basis of the joint state that could not be predicted on the basis of the separate states.... I define a state, λ, formally, as a conditional probability measure, $p_\lambda(x|m)$, assigning probabilities to measured results, x, conditional upon the presence of measurement contexts, m. *With states thus defined, to say that the joint state is wholly determined by the separate states is to say that the joint probability measure is the product of two separate measures* (*ibid.*, p. 226n, our emphasis).

As already suggested, we regard Howard's treatment of states as families of probability measures to be a perfectly sensible way to think about quantum states. As above, in a Bell-type experimental setup, the entire state of the two-particle system can be thought of as being given by a set of functions onto probability measures. We like, moreover, the way Howard lays out the basic intuition of separability—that no information is contained in the joint state that is not already contained in the separate states. The last sentence of the passage, however, the one to which we have added emphasis, stands out to us as a non sequitur. Why, after all, should we equate determination with multiplication, that is, with just the product function?

Surely the joint state is *wholly determined* by the separate states just in case there is *some* function that maps the marginals on each subsystem to joints on the composite system. If it works, the product function would do (so factorizability implies separability). But to restrict our attention to this function as the only possible candidate for separating the two states strikes us as provincial. Any function, plausibly, that maps pairs of separate states onto joint states would make the whole *determinable* by the parts. That, after all, is what a function does—a binary function *determines* an output given any pair of inputs.

Howard's treatment simply equates separability with factorizability. That equation, moreover, is not confined to Howard. Nor is it even confined to discussions of the Bell inequalities. In a recent article on the relation between Leibniz's principle of the identity of indiscernibles and the Pauli exclusion principle, Michela Massimi[21] offers the following definition of separation. Given any composite system, the states of the two component systems are "ontologically separate" just in case:

(1) Each subsystem has definite (though possibly unknown) values of a complete set of compatible observables pertaining to that subsystem alone.
(2) The afore-defined ontologically separate states of the subsystems determine wholly their joint state (*ibid.*, p. 321).

So far so good. She then goes on to say:

> *Condition 2 is known as factorizability*: the joint probability of the expectation values relative to the two subsystems is the product of the probabilities bearing respectively on each subsystem alone (p. 321, our emphasis).

> Factorizability is uncontroversially taken as a separability condition to the extent that all non-factorizable theories involve some kind of non-separability (p. 321n)

We beg to differ. Multiplication is not the only binary function and the failure of factorizability is perfectly compatible with a complete determination of the whole from the separate parts. As we will show, there are local models of experiments in which the Bell inequalities are violated (hence nonfactorizable models) that are indeed separable.

Of course, we understand why someone might think that quantum entanglement rules out separability, which requires that the partial states (the marginals) determine the whole state (the joints). For in an EPR-Bohm experiment, the singlet state of the total system yields a uniform distribution (50:50) as the marginal on the outcome space $\{1, -1\}$ regardless of the angle being measured in either wing. So, for example, the marginals are the same in runs of the experiment with separations of zero degrees between the spin angles measured in the wings (that is, measurements in the same direction) and in runs where the difference between the measured angles is one hundred twenty degrees. But the joint probability functions are different. In the one case, they yield probability 0 for getting identical outcomes in each wing (strict anticorrelation). In the other case, that probability is 3/8. Hence no function maps the marginals of the singlet state to the joints. (Recall that a function determines its result, and so does not map one to many.) Looking at quantum theory this way, it seems that the crucial condition for separability fails; the joint state is not wholly determined by the separate states.

---

[21] "Exclusion Principle and the Identity of Indiscernibles: A Response to Margenau's Argument," *British Journal for the Philosophy of Science*, LII (2001): 303–30.

Before we conclude that, after all, quantum mechanics is not a separable theory, we need to look a little deeper. We need to ask whether the apparent nonseparability of the quantum probabilities is compatible with an underlying separability. Can we, for example, regard the quantum statistics as averages over more basic states, which are themselves separable? This is exactly the sort of issue that the Bell theorem addresses, answering "no." But there the question is whether the conjunction of locality and factorizability can hold at the basic level. Here our concern is with maintaining the conjunction of locality and separability—especially in cases where the Bell inequalities, and hence factorizability, fail.

III. SEPARABILITY WITH LOCALITY: SOME PROOFS

Recall that we are considering correlation experiments violating the Bell inequalities that involve two measurements in each wing (2×2 experiments): call them A_1, A_2 in the A-wing and B_1, B_2 in the B-wing. In what follows, all models preserve locality. That means, for instance, that at each state λ the marginal distribution for A_1 in $A_1 \times B_1$ runs of the experiment is the same as the marginal distribution for A_1 in $A_1 \times B_2$ runs; similarly for the other three pairs. On this assumption, to build a model in which the composite statistics are separable is to define a probabilistic model involving basic states λ that reproduces the quantum statistics as averages over the λ. That model will identify for each state λ, the marginal distributions $p(A_i = x|\lambda)$, $p(B_j = y|\lambda)$ which we understand as the local states in each wing. Then we need to show that there is a single two-place function $F(.\,.)$ such that

$$F[p(A_i = x|\lambda), p(B_j = y|\lambda)] = p_\lambda{}^{AB}(x,y|i,j)$$

for all i, j, x, y, and λ. Thus a single F maps the pairs of marginals at λ to the joints at λ, so that, at every λ, the joint state is wholly determined by the local state. Finally we need to verify that the prescription for local states and joints does yield the quantum probabilities as averages over λ.

Below we take the basic states to be α and β and to begin with we assume these to be equiprobable. Let $a_{\alpha i} = p(A_i = 1|\,\alpha)$. So that $1 - a_{\alpha i} = p(A_i = -1|\alpha)$. Similarly let $b_{\alpha j} = p(B_j = 1|\alpha)$. If the statistics are separable, the A- and B-wing distributions at α are fixed by $a_{\alpha 1}$, $a_{\alpha 2}$ and $b_{\alpha 1}$, $b_{\alpha 2}$ respectively. In the same way we can let $a_{\beta i} = p(A_i = 1|\beta)$

and $b_{\beta j} = p(B_j = 1|\beta)$. Then the A- and B-wing distributions at β are fixed by $a_{\beta 1}$, $a_{\beta 2}$ and $b_{\beta 1}$, $b_{\beta 2}$.

If $a_{\alpha i} + a_{\beta i} = 1$, then

$$1/2[p(A_i = 1|\alpha) + p(A_i = 1|\beta)] = <p(A_i = 1|\lambda)> = 1/2$$

which is the quantum probability that a measurement of A_i yields the result 1. Similarly, if $b_{\alpha j} + b_{\beta j} = 1$, then averaging over the two basic states yields the 50:50 quantum probabilities for each B-wing measurement to have 1 as a result.

Given quantum joint probabilities that violate the Bell inequalities, in order to build a separable model we will assign joint distributions at α and at β that have marginals as above (satisfying locality), and that average to the quantum joints. We will then show that there is a function F mapping the marginals at each of α and β to the joints assigned there, as required for separability. Here is an algorithm for how to do just that for almost all quantum statistics. That is, given "ensemble" joint probabilities, it is an algorithm for finding an α and β and an F that will do the job.

In a 2×2 experiment, quantum theory assigns joint distributions to the four pairs: A_i, B_j for $i,j = 1, 2$. The marginals are all probabilities of 0.5. So if we are given $P^{QM}(A_i = 1, B_j = 1)$, call it $P(A_iB_j)$, then we have an inversion symmetric distribution where

$$P^{QM}(A_i = -1, B_j = 1) = 1/2 - P(A_iB_j) = P^{QM}(A_i = 1, B_j = -1)$$

and

$$P(A_iB_j) = P^{QM}(A_i = -1, B_j = -1)$$

Thus $P(A_iB_j)$ is sufficient to determine the whole $P^{QM}(. .)$.

Then, where λ can be either α or β, assign at λ the joint distribution $p(. . |\lambda)$ defined as follows:

$$p(A_i = 1, B_j = 1|\lambda) = P(A_iB_j) \text{ (that is, the quantum theory joint above)}$$

$$p(A_i = -1, B_j = 1|\lambda) = b_{\lambda j} - P(A_iB_j)$$

$$p(A_i = 1, B_j = -1|\lambda) = a_{\lambda i} - P(A_iB_j)$$

$$p(A_i = -1, B_j = -1|\lambda) = 1 + P(A_iB_j) - a_{\lambda i} - b_{\lambda j}$$

One can check that these return the desired 50:50 marginals, respecting locality, and that averaging over $p(. . |\alpha)$ and $p(. . |\beta)$, so defined, produces exactly the quantum joints provided $a_{\beta i} = 1 - a_{\alpha i}$ and $b_{\beta j} = 1 - b_{\alpha j}$.

Given that each quantum joint satisfies $0 \leq P(A_iB_j) \leq 0.5$, and assuming that the marginals assigned at α and β take values between 0 and 1, the joint distribution $p(.\,.|\lambda)$ defined above requires

$$(1\lambda)\ P(A_iB_j) \leq a_{\lambda i},\ b_{\lambda j}$$

and

$$(2\lambda)\ a_{\lambda i} + b_{\lambda j} \leq 1 + P(A_iB_j)$$

Finally, then, let us choose the marginals. Here is how.

If among the numbers $P(A_iB_j)$ there is a 0 or 1/2, call it $P(A_1B_1)$ and rename the other three accordingly. We are going to assign marginal distributions in the *A*- and *B*-wings using the following scheme (in which it is useful to keep track of the four runs so that we can see that locality is satisfied):

> For run $A_1 \times B_1$ at α : set $p(A_1{=}1|\alpha) = 0.5 = p(B_1{=}1|\alpha)$; that is, set $a_{\alpha 1} = 0.5 = b_{\alpha 1}$.
> For run $A_1 \times B_2$ at α: set $a_{\alpha 1} = 0.5$ (locality!) and set $b_{\alpha 2} = x$, for some $0 \leq x \leq 1$ to be chosen below.
> For run $A_2 \times B_1$ at α: set $a_{\alpha 2} = x$ and $b_{\alpha 1} = 0.5$ (locality).
> For run $A_2 \times B_2$ at α: set $a_{\alpha 2} = x$ (locality) and $b_{\alpha 2} = x$ (locality).

The conditions (1α) and (2α) impose a set of inequalities that x must satisfy. Because of (1β) and (2β) these constraints are multiplied by the requirement that $a_{\beta i} = 1 - a_{\alpha i}$ and $b_{\beta j} = 1 - b_{\alpha j}$. We can summarize all of these inequalities as follows. Call $P(A_1B_2) = t$, $P(A_2B_1) = u$ and $P(A_2B_2) = v$. Then let

$$S = \max[\max(t,\ u\ ,v),\ 0.5 - \min(t,\ u,\ v/2)]$$

and let

$$T = \min[0.5 - \min(t,\ u,\ v/2),\ 1 - \max(t,\ u,\ v/2]$$

Since each joint $(t,\ u,\ v)$ is between zero and 0.5, it is easy to check that $S \leq T$. Then the number x must satisfy the condition that

$$S \leq x \leq T$$

To be specific we can choose $x = S$. Notice that if one (or more) of t, u, v is either 0 or 0.5 then $S = T = 0.5$ and in that case $x = 0.5$. Excluding that case, we now conclude by defining the separating function *F*. Recall that *F* maps the marginals at λ to the joints at λ, for λ either α or β. So let *F* be defined by

$$F[a_{\lambda i}, b_{\lambda j}] = p(A_i{=}1,\ B_j{=}1|\lambda) = P(A_iB_j)$$

We need to check that this leads to no conflicts; that is, that *F* does not take the same input to different outputs. But the input pairs are just $<1/2,1/2>$, $<1/2,x>$, $<x, 1/2>$, $<x, x>$ for α and $<1/2,1/2>$, $<1/2,1-x>$, $<1-x, 1/2>$, $<1-x,1-x>$ for β. Provided $x \neq 1/2$ these are eight distinct pairs, so we can define F on them as indicated without fear of conflict. Obviously when we average over α and β we get back the original quantum joints. Thus, respecting locality, F maps the marginals in the A- and B-wings at λ, which define the states there, to joints at λ on the composite system. Averaged over α and β, these marginals, and the joints they determine via F, yield the single and joint probabilities that quantum theory assigns in the given experiment.

Consider a commonly used illustration for the violation of the Bell inequalities. It is the 2×2 experiment where $P(A_1B_1) = 0$ and $P(A_1B_2) = P(A_2B_1) = P(A_2B_2) = 3/8$. Here the angular separation is 0 between the A_1 setting and the B_1 setting, and all other separations are 120 degrees. Adding the three nonzero joints and subtracting the fourth produces 9/8, whereas the Bell inequality requires a result less than 1. An easy calculation yields that $S = 0.375$, which we choose as our "x" for the separating model. (Here $T = 0.625$.) So the Bell inequality fails but locality and separability both hold in our model. More significantly, consider the experiments at angular separations where the Bell inequality achieves its maximum violation, which is the type of experiment most frequently run and most thoroughly studied. We can take the angles to be 45 degrees between the A_1 setting and the B_1 setting, and 135 degrees for the other three. Then the quantum joint probabilities are $P(A_1B_1) = 0.073$ and $P(A_1B_2) = P(A_2B_1) = P(A_2B_2) = 0.427$. Adding the three big joints and subtracting the fourth little one produces 1.208, which exceeds the Bell limit of 1 by the maximum amount possible for experiments of this type. We can readily model the experiment in a local and separable way, however. S is simply 0.427, which we can take for our "x." (Here $T = 0.573$.)

Our demonstration of separability so far covers all 2×2 experiments where no more than one joint is either 0 or 1/2. That covers a lot of experiments but it does not include the tantalizing Hardy cases that also violate the Bell inequalities and which arise in virtually every interaction between spin-1/2 systems.[22] In the Hardy configuration we have $P(A_1B_1) = 0$, $P(A_1B_2) = 0.5 = P(A_2B_1)$ and $P(A_2B_2) = 0.09$. Our model breaks down here, for $S = T = x = 0.5$ which makes F many-one, and so not a function that determines the joints from the marginals. Indeed, calculations show that no model like ours, where the distributions of the λ states is uniform, can separate the Hardy probabilities. But they can be separated nevertheless. Indeed, we will show how to separate any experiment where $P(A_1B_1) = 0$, $P(A_1B_2) = 0.5 = P(A_2B_1)$, and $P(A_2B_2) = H$ for any number H satisfying $0 < H < 0.5$. (For $H = 0$, the Bell inequalities hold, hence there is a local and factorizable model. At the other extreme, $H = 0.5$ describes an impossible configuration of angles; that is, no quantum experiment produces those numbers.)

---

[22] L. Hardy, "Nonlocality for Two Particles without Inequalities for Almost All Entangled States," *Physical Review Letters*, LXXI (1993), 1665–68.

Our Hardy model is similar to the one just described except that α now gets weight 0.25 and β gets weight 0.75. Moreover the 0 and 0.5 joint probabilities, working together, force a different configuration for the marginals. Again we separate the presentation into disjoint runs to show how locality is respected and also to make it easy to find the pairs of marginals on which F is defined.

For run $A_1 \times B_1$, at α set $a_{\alpha 1} = H$ and set $b_{\alpha 1} = 1-H$.
For run $A_1 \times B_2$, at α set $a_{\alpha 1} = H$ (locality) and set $b_{\alpha 2} = H$.
For run $A_2 \times B_1$, at α set $a_{\alpha 2} = 1-H$ and $b_{\alpha 1} = 1-H$ (locality).
For run $A_2 \times B_2$, at α set $a_{\alpha 2} = 1-H$ (locality) and $b_{\alpha 2} = H$ (locality).

Then at β we define $a_{\beta i}$ and $b_{\beta i}$ to satisfy

$$0.25a_{\alpha i} + 0.75a_{\beta i} = 0.5$$

$$0.25b_{\alpha i} + 0.75b_{\beta i} = 0.5$$

for $i = 1, 2$—so that the marginals averaged over α and β yield the quantum probabilities of 0.5. One can verify that for $0 < H < 0.5$, the marginal probabilities are between 0 and 1 and that the eight pairs of marginals (four at α and four at β) are all distinct. Finally, we assign the joints at α and β as follows:

$$F[a_{\lambda i}, b_{\lambda j}] = p(A_i=1, B_j=1|\lambda) = P(A_iB_j), \text{ for } \lambda = \alpha \text{ or } \beta \text{ and } i = j$$

$$F[a_{\lambda i}, b_{\lambda j}] = p(A_i=1, B_j=1|\lambda) = a_{\lambda i} \text{ for } \lambda = \alpha \text{ or } \beta \text{ and } i \neq j$$

For the case where $i \neq j$ the A and B marginals are the same for each λ, so if we pick one side (above we chose the A side) for the joints there and average over α and β, we get 0.5, since that is precisely how the α and β marginals were made to relate. One needs to check that all the inequalities required by (1λ) and (2λ) hold for these choices of marginals and joints, which they do. So, even the ubiquitous Hardy violations of the Bell inequalities can be modeled in such a way that the local states determine the joints.

IV. DISCUSSION

It is important to keep the quantifiers straight with respect to what we have shown. For each Bell-violating experiment considered above, there exists a function that can be used to determine the joint state from the separate states. This is sufficient to establish that each of those experiments can be given a local, separable model. Nevertheless, we have not proven the following: there exists a single function that can serve as a separations function for all those Bell-type experiments. Such a function would, so to speak, wholly replace the product function as a means of determining joint states from separate states in every single case.

It might be objected that a great deal hinges on this order of quantifiers. After all, one might argue, only if there were a single function could one use the separate states to predict what can be predicted on the basis of the whole state. Is that not, after all, what it means for the separate states *wholly to determine* the joint state?

We note, first of all, that what we have shown so far is already sufficient to refute the lemma according to which, in models of the sort considered above, factorizability was supposed to be a necessary condition for separability. Second, it is important to distinguish the issue of access or predictability from the metaphysical question of separability. Even if there were a universal separating function, we could still be stuck unless that function was computable, or otherwise suitable for making predictions. Looking back to Howard's discussion, we find that he says two inequivalent things about separability:

> The idea is that no information is contained in the joint state that is not already contained in the separate states, or *alternatively*, that no measurement result could be predicted on the basis of the joint state that could not be predicted on the basis of the separate states.[23]

The present context makes it clear, however, that it is one thing to say that the information is contained in the separate states, and quite another to say that one can extract the information and use it to make a prediction, or even that we could know in advance what the separating function will be.

Richard Healy[24] distinguishes these cases, defining what he calls *spatiotemporal separability* as follows:

---

[23] "Holism, Separability, and the Metaphysical Implications of the Bell Experiments," p. 226n, our emphasis.

[24] "Holism and Nonseparability," this JOURNAL, LXXXVIII, 8 (August 1991): 393–421.

> The qualitative, intrinsic physical properties of a compound system are *supervenient* on the qualitative, intrinsic physical properties of its spatially separated component systems together with the spatial relations among these component systems (*ibid.*, p. 410).

The language of supervenience marks a criterion other than that of predictability. In fact, Healy contrasts the relevant principle with what he calls *explanatory principles of holism*—principles, which depend on predictability for their application.

Since the original inspiration for the concept of separability came from Einstein, perhaps it would be helpful to go back and look at what Einstein himself had in mind. He wrote:

> [Without a principle of separation] physical thought in the sense familiar to us would not be possible. Nor does one see *how physical laws could be formulated and tested* without such a clean separation.[25]

If Einstein's concern is for the testability of laws, then the concern would likely be along the following lines. We make measurements in spatially local contexts. If we are going to have a theory that we can test, then we had better be able to use the information that we gather locally to build up what we need to know in order to tell what a given theory predicts about the world. In other words, it had better be the case that everything that is predictively relevant about a system can be humanly put together from locally separate states of affairs. For if there are theoretically/predictively important aspects of a system that we cannot come to know by making local measurements, then we will not be able to know what outcome a theory predicts for a given state of affairs in important cases. Hence, we would not be able to test such a theory.

If the above reconstruction of Einstein's thoughts is roughly correct, then it should be clear that Einstein had in mind, after all, something over and above metaphysical separability. The concern that Einstein evinces for the testability of physical laws forces this conclusion. By itself, separability does not address Einstein's concern about how quantum mechanics could possibly work in practice. For that one needs a (metaphysical) separability that is also operational. That is, to satisfy Einstein, one needs more than a separating function for every experiment, or even a universal function. What one needs are humanly identifiable, computable functions.

But if Einstein advocated a metaphysical separability that was also operational, because he believed that this was a necessary condition for the testability of a theory, and if he was right, then we are left with a strange conundrum. For quantum mechanics does not appear to be even metaphysically separable; forget about the operational criterion. So the conundrum is this: If, according to Einstein, separability is the *sine qua non* of a testable theory, then how does quantum mechanics work? That is to say, if quantum mechanics is a theory in which the principle of separation actually fails, how does it come not only to have been tested, but to have passed with flying colors? How does quantum mechanics do what Einstein denies can be done by any theory that violates the principle of separability?

---

[25] Quoted in Howard, "Holism, Separability, and the Metaphysical Implications of the Bell Experiments," p. 233.

Of course, we could conclude that Einstein was just wrong here. But assuming he was on the right track, a good way to approach this question is to re-examine the notion of "state" in the quantum theory. Thus far, we have taken a state to be a family of probability distributions. But there is an ambiguity here familiar to all students of mathematics: the ambiguity between the values of a function (what Gottlob Frege called "the course of values") and the function itself. In our models, and in discussing the quantum probabilities as states, we took the local states to be the bare distributions, which is to say the *values of the mapping*—from detector settings (or angles) to probability distributions on $\{1, -1\}$. But if we say that the local state is not the value of this mapping but the mapping itself, then we have a fresh way of regarding the quantum theory. Perhaps the best way to bring this out is to display the local states as a set of pairs $<z, P_z(.)>$, where z is an angle corresponding to one of the detector settings in a given wing and $P_z(.)$ is the marginal distribution on $\{1, -1\}$ at z. Then separability asks whether there is a function $F[.\ .]$ such that for all states $<a, P_a(.)>$ in the A-wing and $<b, P_b(.)>$ in the B-wing $F[<a, P_a(.)>, <b, P_b(.)>]$ is the joint state of the composite system. Put in this more perspicuous way, the answer is clear. There is such a universal separating function; namely,

$$F[<a, P_a(.)>, <b, P_b(.)>] = <<a, b>, 1/2\sin^2[(a-b)/2]>$$

the joint probability function in the singlet state that assigns to the angle a in the A-wing and b in the B-wing the probability $1/2\sin^2[(a-b)/2]$ for spin-up in both directions. This is, moreover, a simple function, easy to compute and easy to use for predictions. Notice that this F depends only on the domain of the marginal distributions (the angles a and b), not on the range. It must do so since all the distributions $P_a(.)$ and $P_b(.)$ are the same. Still, on the present, more careful conception of state, what the function F achieves is *wholly to determine* the state of the whole system from the state of the parts. Einstein's concerns are addressed in the best possible way. There is a universal separating function and one entirely suitable for predicting and testing. It would be hard to ask for more by way of quantum separability, and it would be hard to imagine a stronger defeat for the "lemma" that separability implies factorizability.

ERIC WINSBERG

University of South Florida/Tampa

ARTHUR FINE

University of Washington/Seattle

BOOK REVIEWS

Context and Content. ROBERT STALNAKER. New York: Oxford University Press, 1999. 283 p. Paper $19.95.*

This welcome collection brings together fourteen important papers written by Robert Stalnaker over the last thirty years (one appearing here for the first time). They focus on a variety of overlapping topics dealing with the intentionality of thought and speech, and the interaction of context with what we think and say. Also included is a wonderful introductory essay, offering an especially clear and helpful overview of the issues discussed in the various papers, identifying background assumptions and overlapping themes.

These papers (together with his 1985 book, *Inquiry*[1]) record Stalnaker's efforts over three decades to develop and explore an intricate and compelling framework for thinking about issues of intentionality. Among the central commitments of this framework are the following: intentional states like belief, desire, and intention are to be explained in terms of relations that their subjects bear to propositions; linguistic intentionality is to be explained in terms of the intentionality of mental states; propositional content is conceptually independent of any means of representing it; and the way fundamental intentional states relate their subjects to propositions is to be explained naturalistically in terms of the causal and counterfactual constructions that best characterize the different states. As Stalnaker develops them, these features are, in turn, supposed to underwrite a possible-worlds analysis of propositions, according to which propositions are represented as functions from possible worlds to truth values—or (more or less equivalently) as sets of possible worlds.

The essays in the present collection are an indispensable complement to the discussion found in *Inquiry*, which focused largely on detailing and defending the connection between the naturalistic picture of intentionality that Stalnaker favors and the possible-worlds analysis of propositions. They offer additional insight into the nature of the framework, and, perhaps most importantly, they attempt to provide detailed applications of the framework to specific philosophical problems. Our

---

*I am grateful to Robert Stalnaker for his comments

[1] Cambridge: MIT, 1985.

0022-362X/03/0001/98–108

assessment of this intricate package of ideas will largely depend on how successful we judge these proposed applications to be. In a range of important cases, our judgment will depend on how plausibly Stalnaker's ingenious deployment of certain devices borrowed from two-dimensional modal logic overcomes or mitigates familiar counterintuitive consequences of adopting the possible-worlds analysis of propositions—about which more below.

The essays are grouped thematically into four sections. Section I, "Representing Contexts," contains five papers: "Pragmatics," "Pragmatic Presuppositions," "Indicative Conditionals," "Assertion," and "On the Representation of Context." They concern the interpretation of speech acts, focusing on how what is said depends on context, and how context itself is changed by what is said.[2] It is in order to illuminate some of these issues that Stalnaker first deploys the formal apparatus of two-dimensional modal logic. In the remaining papers, the primary focus shifts away from speech to thought. In Section II, "Attributing Attitudes," the three essays—"Semantics for Belief," "Indexical Belief," and "Belief and Context"—are transitional, addressing certain puzzles that arise in connection with our use of language to ascribe thoughts. Relying heavily on the two-dimensional apparatus introduced in the previous section, Stalnaker's main aim here is to show both that and how contextual factors play an essential role in determining which proposition is denoted by the *that-clause* in belief reports. Section III, "Externalism," contains "On What's in the Head," "Narrow Content," and "Twin Earth Revisited." These essays articulate and defend the kind of content externalism implicit in Stalnaker's naturalistic approach to intentionality. He persuasively demonstrates how difficult it is to make sense of anything worth calling narrow *content* and argues that, contrary to what is often supposed by others, the two-dimensional apparatus previously deployed offers no help here. The essays in the final section, "Form and Content"—which includes "Mental Content and Linguistic Form," "The Problem of Logical Omniscience, I," and, appearing here for the first time, "The Problem of Logical Omniscience, II"—are more direct attempts to explore and defend those features of his overall framework that attract him to the possible-worlds analysis of propositions. As Stalnaker himself admits, these papers do not attempt to offer a solution to the problem of logical omniscience; rather, they attempt to clarify how best to understand the problem and to show that it is a deeper problem than is usually assumed, one not obviously avoided by adopting a more fine-grained analysis of propositions.

---

[2] Stalnaker proposes that we think of a context as a body of information presupposed to be available to the participants in the conversation. This proposal is motivated, on the one hand, by the plausible idea that a speaker, in aiming to be understood, must suppose that the information relevant to determining the content of his utterance is available to his audience and, on the other hand, by the assumption that a speaker normally aims to add to or alter a body of information that is presumed to be shared by the participants in the speech situation.

For many philosophers, the coarse-grained identity conditions for propositions offered by the possible-worlds analysis does not fit well with our ordinary judgments about when speakers say or believe the same or different things. For example, it has seemed obvious to many that one can assert or believe that

(1) Hesperus is Venus

while failing to assert or believe that

(2) Phosphorus is Venus.

—indeed, that one can believe that (1) even while believing that

(3) Phosphorus is not Venus.

or that

(4) Phosphorus is Mars.

But if (a) assertion and belief are (binary) relations between subjects and propositions, with the latter viewed as sets of possible worlds, and if (b) the relevant propositions in these cases are the ones that a correct descriptive semantics for our language assigns the sentences (1)–(4), then this will be impossible.[3] For, in cases of the first sort, the sentences, (1) and (2), will be assigned the same (necessary) proposition,[4] and no plausible view committed to (a) can allow that someone both believes and does not believe the same proposition. Moreover, according to (a), if I believe a proposition p, and p logically/metaphysically entails the proposition q (that is, $p \subset q$) then, I also believe the proposition q. This is just the problem of logical omniscience mentioned earlier. Thus, in cases of the second sort, where one of the propositions believed is impossible, assumptions (a) and (b) commit us to supposing that our subject believes every proposition, which is surely unacceptable.

---

[3] In "Reference and Necessity," in Bob Hale and Crispin Wright, eds., *A Companion to the Philosophy of Language* (New York: Cambridge, 1997), pp. 534–54, Stalnaker distinguishes between what he calls *descriptive* and *foundational* semantic theories. Roughly, a descriptive semantic theory for a language assigns semantic values to all the expressions in the language, while a foundational semantics specifies the facts in virtue of which a language has a particular descriptive semantics.

[4] I am following Stalnaker here in assuming that ordinary proper names are rigid designators.

Stalnaker's attempts to avoid problems like these lean heavily on his innovative deployment of the two-dimensional apparatus previously mentioned. According to Stalnaker, we are mistaken if we assume that what is said on the occasion of asserting a sentence *S* is *always* the very proposition that a correct descriptive semantics for the relevant language will assign to *S*. Rather, the proposition expressed by an assertoric utterance is context dependent—context dependent in a way that goes beyond simple indexicality. Similarly, we are mistaken if we assume that the proposition denoted by the that-clause of a belief ascription must always be the same proposition as the one that the descriptive semantics assigns to the content sentence occurring therein. Consequently, assumption (b) above does not always hold. What the two-dimensional apparatus enables Stalnaker to do is to identify, in problematic cases, a candidate proposition for what gets said or believed that is different from the one normally assigned to the relevant sentence, and that avoids the difficulties so far mentioned.[5]

The central notion deployed here is that of a *propositional concept*—a function from possible worlds to propositions. Supposing that *a*, *b*, and *c* exhaust the possible worlds with which we are concerned, a propositional concept can be represented by a two-dimensional matrix like the following:

$$\begin{array}{cccc} & a & b & c \\ a & \mathrm{T} & \mathrm{T} & \mathrm{T} \\ b & \mathrm{F} & \mathrm{F} & \mathrm{T} \\ c & \mathrm{F} & \mathrm{T} & \mathrm{T} \end{array}$$

This matrix helps us to see that a propositional concept can be used to determine two different propositions. On the one hand, each row of the matrix represents a proposition that is the value of the propositional function for each possible world that serves as an argument. Stalnaker calls these *horizontal propositions*. But there is also the proposition that is true at a world *w* just in case the proposition that is the value of the propositional concept for *w* is true at *w*. Stalnaker calls this the *diagonal proposition*, since it can be read off from the diagonal of the matrix.

---

[5] The devices from two-dimensional modal logic that Stalnaker exploits have attracted considerable attention recently, but have been used by different philosophers in quite different ways for quite different purposes. Specifically, the interpretation and uses to which Stalnaker puts them are quite different from the uses to which they are put by David Chalmers in *The Conscious Mind* (New York: Oxford, 1996), especially pp. 56–65, and by Frank Jackson in *From Metaphysics to Ethics: A Defense of Conceptual Analysis* (New York: Oxford, 1998), especially pp. 46–86.

Stalnaker suggests that we can use a propositional concept to represent a person's epistemic situation with respect to specific token utterances. This will be especially useful, he claims, where someone (whether speaker or audience) has incomplete knowledge of the facts relevant to determining the semantic content of some utterance. Assuming that the relevant token utterance occurs in each of the possible worlds in what Stalnaker calls the subject's *context set*—the set of possible worlds compatible with the information she presupposes to be available to the participants in the conversation—we can use a propositional concept, defined on these worlds, to represent such a person's epistemic situation with respect to what is said by the utterance. The horizontal propositions will represent what is said "according to the normal semantic rules" (14) that are operative in the relevant possible worlds (those belonging to the context set). Since, presumably, it is a contingent matter which proposition is expressed by the counterpart utterance in each of these worlds, the horizontal propositions will not all be the same.[6] The diagonal proposition will represent the information the subject would possess if she came to know that the token utterance expressed a truth, even if she did not know what truth it expressed. What Stalnaker proposes is that in at least some cases—specifically, problematic cases of the sort noted previously—we identify the proposition expressed by the speaker's utterance, or denoted by the that-clause in a belief ascription, with the diagonal proposition determined by the appropriate propositional concept.[7]

Consider, for example, a case in which someone assertorically utters a necessarily true or necessarily false sentence such as our

---

[6] My use of "counterpart" here need not be taken counterpart-theoretically. If not, then what I am calling the counterpart of a particular actual utterance will be that very token utterance in the relevant possible world.

[7] Stalnaker deploys this apparatus to clarify what is going on in a variety of philosophically puzzling cases—assertions involving referential uses of definite descriptions, negative existential claims, Kripke-style cases of necessary a posteriori and contingent a priori statements, as well as cases of mistaken identity. I shall, however, confine my attention to his use of it to avoid the sorts of difficulties noted in the text. I shall not address the question whether the success or failure of Stalnaker's deployment of it there has any bearing on the success or failure of its deployment in these other puzzling cases, though in some cases the implications should be clear.

(1) Hesperus is Venus.

or

(4) Phosphorus is Mars.

Since, for Stalnaker, a speaker's aim in making an assertion is to reduce the context set—to eliminate from the possibilities that are left open by the mutual presuppositions of the conversational participants those that, according to the speaker, are not the way the world is—a puzzle arises about how we are to understand what is said when she assertorically utters such sentences. After all, the proposition semantically associated with (1) in the actual world, being necessary, is not incompatible with any possibility that the conversational participants mutually presuppose. To aim to eliminate the possible worlds incompatible with (1) from the context set would be pointless, for there are none. And since the proposition semantically associated with (4) is necessarily false, it would be absurd to attempt to eliminate the possible worlds incompatible with (4) from the context set, for that would reduce the context set to the null set. Given the aim of assertion, appeal to broadly Gricean conversational principles will require, as Stalnaker sees it, that the normal semantic rules governing the asserted sentence be overridden in such cases. The horizontal proposition determined by the relevant propositional concept for the actual world, if the actual world belongs to the context set, will *not* count as what is said. Stalnaker suggests that in such cases we should identify what is said with the relevant diagonal proposition. Since this will be a contingent proposition, our problem disappears.

While this is an ingenious and suggestive proposal, our ultimate assessment of its promise and plausibility rests on exactly how we are to understand its content. And this, I suggest, is not altogether clear. First of all, since Stalnaker does not claim that what is said on the occasion of an assertoric utterance is always the diagonal proposition—though, presumably, there will always be such a proposition—it would be nice to have a principled account of just when the diagonal proposition is and when it is not what is said. It is implausible to suppose that we should diagonalize, so to speak, whenever the speaker presupposes false semantically relevant information—assuming for the moment that we have a clear sense of what counts as semantically relevant information. Nor, obviously, will it be very satisfying to be told to diagonalize whenever the proposition assigned by the normal semantics leads to trouble. Stalnaker, of course, does not rely on anything quite so ad hoc. As we saw, he seeks to motivate the proposal by appealing to a suggestion about the aim of assertion and to certain Gricean conversational principles.[8] But, as Stalnaker would doubtless be the first to admit, both his gloss on the constitutive aims of assertion and the conversational principles to which he appeals, while perhaps pointing in the right direction, are insufficiently detailed to provide a principled and sufficiently general account of when we should identify what is said with the diagonal proposition.

Another respect in which Stalnaker's suggestion is unclear concerns how we are to understand the diagonal proposition that he identifies with what is said in these problematic cases. How, for example, might we intuitively express this proposition? The informal glosses that Stalnaker offers make it unclear whether this will be some variety of metalinguistic proposition ('Phosphorus' refers to the same thing as 'Venus'; 'Phosphorus' refers to the brightest astronomical body seen...) or a fully object-level proposition (the first astronomical object seen in the morning is the same as the brightest astronomical body seen...).

This uncertainty is related to another and perhaps more troubling respect in which Stalnaker's proposal is unclear. For any given utterance, a wide variety of propositional concepts can be defined. Which is the relevant one? If I understand Stalnaker correctly, the relevant propositional concept is the one that delivers as values, for each possible world in the context set, the proposition that constitutes what, according to the normal semantic rules for that world, is said by whatever utterance in that world is the counterpart of the actual utterance. But this characterization raises a number of interrelated questions. First, what are the criteria for being a counterpart utterance of the original? Must a counterpart utterance occur in a counterpart context? In particular, are there any substantive *semantic* constraints on what is to count as a counterpart utterance? If so, what are they and what determines them? Stalnaker, recall, allows that the semantic rules operative in alternative worlds may differ from those in the actual world. But how different can they be and in what respects, such that an utterance can so much as count as a candidate counterpart to the original? All of this, of course, is just another way of asking for more guidance in determining which worlds belong in the context set.

---

[8] It is worth noting, however, that it is only on the assumption that propositions are given a possible-worlds analysis that these Gricean principles engage with Stalnaker's suggestion about the aim of assertion in such a way as to underwrite the claim that the proposition expressed by our subject's utterance of (1) or (4) cannot plausibly be the proposition determined by the normal meaning of the sentence uttered. Apart from this assumption, there is no obvious way in which the Gricean principles combined with Stalnaker's suggestion about the aim of assertion—*insofar as it can be expressed in a theory-neutral way*—force us to suppose that what is said cannot plausibly be the proposition determined by the normal semantics.

Presumably, Stalnaker will say that the semantic rules can be as different as is consistent with the information the subject supposes to be mutually available to the participants in the conversation. This will follow from the idea that the propositional concept is defined relative to the context set. But it is far from clear what may reasonably be taken to count as mutually presupposed semantic or semantically relevant information. Does the relevant information concern the descriptive semantics of the language, or the foundational semantics, or both?[9] For example, is it presupposed that 'Phosphorus' refers to Phosphorus, 'Hesperus' to Hesperus, 'Venus' to Venus, and 'Mars' to Mars? While for many (myself included) this will seem like a natural suggestion, it would lead to obvious problems for Stalnaker's proposed solution. But, apart from the fact that it leads to trouble, what good reasons are there for denying that this information is part of what is presupposed? Perhaps, instead, Stalnaker would have us focus on the speaker's presuppositions concerning foundational semantic information. But as we shall see, there is potential for trouble here too. In any case, it is at best unclear what to say if the speaker presupposes a mistaken foundational semantics (as many philosophers believe other philosophers to do), or makes no substantive presuppositions whatsoever about the foundational semantics. In short, until we have a better sense of the sorts of semantic or semantically relevant information that bear on what is to count as a counterpart utterance, and thus on which worlds belong in the context set, we shall have no clear view of how to define or interpret the relevant propositional concept and, so, no way of identifying the relevant diagonal proposition.

So far, these considerations may merely suggest that Stalnaker owes us more details if we are adequately to assess the promise of his two-dimensionalist strategy for dealing with problematic cases. But they point, I believe, to more serious difficulties that emerge most clearly with respect to Stalnaker's proposals concerning belief ascriptions. Stalnaker is clear that in order to extend the diagonalization strategy to belief ascriptions (where the issue, recall, is how to identify the proposition referred to by the that-clause), we shall need to define a propositional concept not only for the context set—the possible worlds compatible with the speaker's presuppositions—but also for the set of possible worlds that are presupposed to be compatible with what the subject of the attribution believes—that is, her doxastic possibilities. After all, what the subject believes or is taken to believe may be quite different from, indeed, incompatible with what the speaker presupposes. The subject may, for example, be unaware of the existence of the speaker and may not even speak the same language. So, in particular, she may have no beliefs about the speaker's utterance. Stalnaker calls the set of worlds presupposed to be compatible with what the subject believes the *derived context*.

---

[9] See footnote 3 above.

To see how this is supposed to help us in the case of problematic belief reports, consider a situation in which we want to count as true Bob's assertion of the following:

(5) *S* believes that Phosphorus is Mars.

The content sentence here is, of course, our (4). As already noted, it would be implausible, given the possible-worlds analysis of propositions, to suppose that Bob is asserting that the proposition *S* believes is the very proposition that the correct descriptive semantics (in the actual world) assigns to (4)—namely, the empty set. So, if we are to respect our assumption that Bob's assertion is true, we must suppose that a different proposition is expressed by the speaker's utterance of (4), and that it is this other proposition that is denoted by the that-clause containing it. According to Stalnaker, to determine which proposition this is, we first identify all the possible worlds in the derived context in which a counterpart of Bob's utterance of (4) exists. We then determine what proposition would be expressed by that utterance according to the normal semantic rules for that world. This will define the relevant propositional concept. Stalnaker's proposal is that the diagonal proposition determined by this propositional concept is the proposition expressed by Robert's utterance of (4)—in the context of his asserting (5). And so, this diagonal proposition is the one Bob truly asserts that *S* believes.

Here, though, our earlier worries about how exactly we are supposed to define the relevant propositional concepts lead immediately to serious difficulties. Suppose that *S* is not an English speaker and has no opinions about the normal semantic rules for English. Consequently, the range of possible semantic rules that determine the proposition expressed by the counterpart to Bob's utterance of (4) is virtually unconstrained. In such a case, we have every reason to expect that there will be at least one possible world w^*—that is compatible with what *S* believes, and in which the semantic rules determine that the proposition expressed by the counterpart of Bob's utterance of (4) is necessarily false. In this case, the diagonal proposition will be false at w^*—that is, it will be false at one of S's doxastic possibilities. But then, contrary to our original assumption, it will be false that S believes the diagonal proposition. For a possible world w represents a doxastic possibility for a subject just in case, for every proposition p that she believes, $w \in p$. But w^* is not a member of the diagonal proposition.

Unless we have some non-ad hoc way of excluding from the derived context those worlds whose semantics assign a necessarily false proposition (or any other proposition that is false at one of the subject's doxastic possibilities) to the counterpart utterance of the content sentence in the original belief report, Stalnaker's strategy will force us to regard far too many belief reports as false. Will it help here to appeal to the subject's foundational semantic beliefs? Perhaps in some cases, but not as a general rule, for there seem to be no compelling reasons to think that speakers must presuppose that their subjects have sufficiently correct or detailed enough beliefs about such matters to preclude the problematic worlds' belonging to the derive context.[10] The problem arises because, as Stalnaker deploys the two-dimensional apparatus, any constraints on the semantics in the worlds for which the propositional concept is defined must derive from (what is presupposed about) the doxastic situation of the subject. But it is not always plausible to suppose that the subjects of our belief reports have the kinds of beliefs that will exclude the problematic worlds from their doxastic possibilities. Perhaps principled reasons can be found for introducing *extra*-doxastic constraints on the set of worlds for which the relevant propositional concept is to be defined, constraints that successfully exclude the problematic worlds and yield a diagonal proposition that is a plausible candidate for what, in our problematic cases, a speaker asserts is believed by the subject. It is difficult, however, to see what these additional constraints might be or how they might be motivated. And in any case, this sort of move would seem to involve a substantial shift in how Stalnaker would have us interpret the resulting propositional concepts.

---

[10] These considerations may also raise various difficulties for Stalnaker's account of assertion. For example, unless we have good reasons for supposing that the semantics operative in those members of the context set that are compatible with the speaker's beliefs will never assign a necessary false proposition to the relevant utterance, then it may well turn out that the proposition asserted, when this is supposed to be the diagonal proposition, cannot be a proposition the speaker believes—even in cases where the subject is not intending to say something he believes to be false.

Needless to say, I have not been able to address all of the interesting and important ideas that Stalnaker develops and defends in the various papers in the volume under review. I have focused instead on some concerns about the adequacy and promise of Stalnaker's two-dimensionalist strategy for avoiding some of the familiar difficulties that beset possible-worlds analyses of propositional content. I think these concerns suffice to raise a genuine challenge to the framework for understanding intentionality that Stalnaker would have us adopt. I do not claim that there might not be possible alternative strategies, consistent with the general framework, which might more successfully deal with these difficulties. While I am not sanguine about the prospects of finding such an alternative, this wonderful collection of papers, always subtle and ingenious, is the best evidence that at least one philosopher is very well positioned to identify and develop just such a strategy. Philosophers, linguists, and cognitive scientists interested in the issues discussed in these papers cannot help but find this collection challenging, provocative, and full of important insights.

WILLIAM W. TASCHEK

The Ohio State University

NOTES AND NEWS

The Editors regret the death of John Rawls, James Bryant Conant University Professor of Philosophy at Harvard University. He died on November 24, 2002, at the age of 81. Professor Rawls enjoyed a long and distinguished career as a political philosopher and a powerful advocate of the liberal tradition.

Rawls was born in Baltimore in 1921. After earning a B.A. from Princeton University, he served as a combat infantryman in the South Pacific from 1943–45. He then returned to Princeton the following year, and he received his doctorate in philosophy in 1950. Rawls was an Instructor at Princeton from 1950–52, and received a Fulbright Fellowship to Oxford University from 1952–53. He was Assistant and Associate Professor at Cornell University from 1953–59, and Professor of Philosophy at the Massachusetts Institute of Technology from 1960–62. He joined the philosophy department at Harvard in 1962, becoming Conant University Professor in 1979.

Professor Rawls's distinguished career occasioned numerous awards and honors. He served as President of both the American Philosophical Association in 1974 and the American Association of Political and Legal Philosophy from 1970–72; he was a member of the American Academy of Arts and Sciences, the American Philosophical Society, the British Academy, and the Norwegian Academy of Sciences. Rawls won a National Humanities Medal in 1999, with a citation for "helping women enter the ranks of a male-dominated field," and was offered numerous honorary degrees, accepting them only from universities with which he was associated (Oxford, Princeton, Harvard).

His published work includes *A Theory of Justice* (1971), *Political Liberalism* (1993), *Collected Papers* (1999), *The Law of Peoples; with, The Idea of Public Reason Revisited* (1999), *Lectures on the History of Moral Philosophy* (2000), and *Justice as Fairness: A Restatement* (2001). Professor Rawls's was published several times in this *Journal*'s pages, including an early version of "Justice as Fairness" in 1957, his 1980 Dewey Lectures, "Kantian Constructivism in Moral Theory," and a 1995 exchange with Professor Jürgen Habermas.

Professor Rawls' best known book, the widely read *A Theory of Justice*, was nominated for a National Book Award. In it, building upon the theories of Hobbes and others, Rawls laid out his own idea of a "well ordered society," one that is founded upon a social contract. In order to best protect the interests of each and every member of a society, he posits, members must, while choosing the principles of justice by which they wish to be governed, wear a "veil of ignorance" about their exact stations in life. This hypothetical situation, or "original position," is designed ultimately to minimize social inequality and to protect basic rights and opportunities for all. These famous arguments underwent constant revision over time as Rawls focused increasingly on the questions raised by pluralism and responded to the sizeable literature that his work inspired. Given the undeniable influence of his work on the profession, Professor Rawls has been considered by many to be the most important political philosopher of the second half of the twentieth century.

The Editors regret the death of Richard C. Jeffrey, Professor Emeritus of Philosophy at Princeton University. He died on November 9, 2002, at the age of 76. Known foremost as an expert in probability and decision theory, Jeffrey also taught logic and the philosophy of science.

Professor Jeffrey earned his Ph.D. in philosophy at Princeton in 1957. A World War II Navy veteran, he earned an M.A. from the University of Chicago in 1952. He was a research engineer, a Fulbright Scholar, and an Assistant Professor of electrical engineering before starting his career teaching philosophy at Stanford University in 1959. Jeffrey served as a visiting faculty member at the Institute for Advanced Study, and at Princeton in 1963–64. He also taught at the City College of New York and at the University of Pennsylvania before returning to Princeton as Professor of Philosophy in 1974. In recent years he spent part of his time in Princeton, and part as Visiting Distinguished Professor of Logic and Philosophy at the University of California, Irvine.

Jeffrey's Princeton thesis, "Contributions to the Theory of Inductive Probability," explored the treatment of uncertain evidence that later became famous as Jeffrey Conditioning. Jeffrey's subsequent book, *The Logic of Decision* (1965), presented his new theory of decision-making under uncertainty and of probable knowledge, in which both probability and value attach to each proposition describing the world. He further developed these themes, which formed the core of a philosophy he called "Radical Probabilism," in a series of articles and *Probability and the Art of Judgment* (1992). His textbook, *Computability and Logic* (with George Boolos, 1974), bridged the gap between general books on logic and treatises written for mathematicians. He published numerous articles and reviews over the years in this *Journal*.

Jeffrey was elected to the American Academy of Arts and Sciences in 1994, and he served as president of the Philosophy of Science Association from 1999 to 2000. He was the recipient of numerous international awards and medals.

THE JOURNAL OF

PHILOSOPHY

SUBSCRIPTIONS

2003

Students $20.00
Individuals $35.00
Institutions $75.00
Postage outside the U.S. $15.00

SPECIAL OFFER

one complimentary copy of any of the following special issues with each new individual subscription

2000	*Causation*	David Lewis, Ned Hall, John Collins, L.A. Paul, Jonathan Schaffer
1998	*Having the World in View: Sellars, Kant, and Intentionality*	John McDowell
1995	*Reconciliation through the Public Use of Reason*	Jürgen Habermas
	Reply to Habermas	John Rawls
1994	*Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind*	Hilary Putnam
1990	*The Structure and Content of Truth*	Donald Davidson
1985	*Well-Being, Agency, and Freedom*	Amartya Sen
1980	*Kantian Constructivism in Moral Theory*	John Rawls
1977	*Teleology Revisited*	Ernest Nagel
1968	*Ontological Relativity*	W.V. Quine

All issues above are also available for $8.00 plus $3.00 for shipping/handling ($5.00 for international and Canadian orders). Postage for each added issue, $1.00. Send check or money order to:

Journal of Philosophy
709 Philosophy Hall
1150 Amsterdam Avenue, MC 4972
Columbia University
New York, NY 10027
(212) 666-4419
www.journalofphilosophy.org

Social Philosophy & Policy

Editor:
Ellen Frankel Paul,
Bowling Green State University, USA

...is an interdisciplinary journal with an emphasis on the philosophical underpinnings of enduring social policy debates. The issues are thematic in format, examining a specific area of concern with contributions from scholars in philosophy, economics, political science and law. While not primarily a journal of policy prescriptions, some articles in each issue will typically connect theory with practice. The 2003 issues will be: 'After Socialism' and 'Autonomy'.

Social Philosophy and Policy is far and away the best English-language journal devoted to issues of moral and political philosophy that have a direct and important bearing on policy. It has no close rival.
-Allen Buchanan, Grainger Prof. of Business Ethics, University of Wisconsin, Madison

semi-annual
vol 20, 2003

Inst. (print + e)
$128.00
(0265-0525)

Inst. (e only)
$113.00
(1471-6437)

Indiv. (print only)
$44.00
(0265-0525)

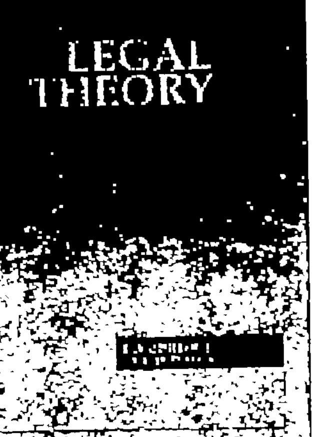
email: journals_subscriptions@cup.org
forty west twentieth street
new york, ny 10011-4211
tel: 800 872-7423
fax: 914 937 4712

online

CAMBRIDGE
UNIVERSITY PRESS

LEGAL THEORY

Legal Theory draws contributions not only from academic law, but from a wide range of related disciplines in the humanities and social sciences, including philosophy, political science, economics, history, and sociology. Topics covered fall mainly into the broad categories of analytical and normative jurisprudence, doctrinal theory, policy analyses of legal doctrines and critical theories of law.



Editors:
Jules L. Coleman
Yale Law School
Larry Alexander
Univ. of San Diego Law School

subscribe, quarterly, volume 9 in 2003
institutions, *print + e (1352-3252)*: $163.00
institutions, *e-only (1469-8048)*: $148.00
individuals, *print only (1352-3252)*: $65.00

Legal Theory *is especially relevant for:*

- lawyers
- academic law practitioners
- legal organisations
- experts in philosophy, political science, history and sociology
- government departments
- students and teachers of law
- judges

40 west 20th street
new york, ny 10011-4211
tel: (800) 872-7423
fax: (914) 937-4712

CAMBRIDGE
UNIVERSITY PRESS

web: http://www.cambridge.org

New and Noteworthy

Natural Rights and the Right to Choose
Hadley Arkes
0-521-81218-6 Hardback $28.00

On the Trinity
Augustine
Gareth Matthews, Editor
Cambridge Texts in the History of Philosophy
0-521-79231-2 Hardback $55.00
0-521-79665-2 Paperback $20.00

The Improvisation of Musical Dialogue
A Phenomenology of Music
Bruce Ellis Benson
0-521-81093-0 Hardback $53.00
0-521-00932-4 Paperback $19.00

Classic and Romantic German Aesthetics
J.M. Bernstein, Editor
Cambridge Texts in the History of Philosophy
0-521-80639-9 Hardback $60.00
0-521-00111-0 Paperback $22.00

Nietzsche: Writings from the Late Notebooks
Rüdiger Bittner, Editor
Cambridge Texts in the History of Philosophy
0-521-80405-1 Hardback $50.00
0-521-00887-5 Paperback $18.00

The Cambridge Companion to Simone de Beauvoir
Claudia Card, Editor
Cambridge Companions to Philosophy
0-521-79096-4 Hardback $60.00
0-521-79429-3 Paperback $22.00

Heidegger's Analytic
Interpretation, Discourse, and Authenticity in Being and Time
Taylor Carman
Modern European Philosophy
0-521-82045-6 Hardback $60.00

Disowning Knowledge
In Seven Plays of Shakespeare
Updated Edition
Stanley Cavell
0-521-82189-4 Hardback $60.00*
0-521-52920-4 Paperback $23.00*

Must We Mean What We Say?
A Book of Essays
Updated Edition
Stanley Cavell
0-521-82188-6 Hardback $70.00
0-521-52919-0 Paperback $25.00

Stanley Cavell
Richard Eldridge, Editor
Contemporary Philosophy in Focus
0-521-77025-4 Hardback $60.00
0-521-77972-3 Paperback $20.00

Radical Interpretation in Religion
Nancy K. Frankenberry, Editor
0-521-81686-6 Hardback $65.00
0-521-01705-X Paperback $24.00

The Cambridge History of Seventeenth-Century Philosophy
Volumes 1 and 2
Daniel Garber and
Michael Ayers, Editors
0-521-53180-2
Two-Volume Paperback Set $75.00*

from Cambridge

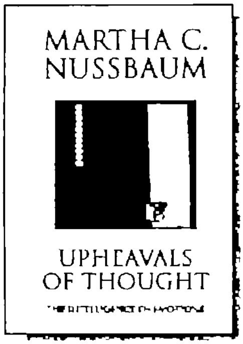
Kierkegaard
A Biography
Alastair Hannay
0-521-53181-0 Paperback $27.00*

Schleiermacher: Lectures on Philosophical Ethics
Robert B. Louden, Editor
Cambridge Texts in the History of Philosophy
0-521-80982-7 Hardback $60.00
0-521-00767-4 Paperback $21.00

Realism and Christian Faith
God, Grammar, and Meaning
Andrew Moore
0-521-81109-0 Hardback $65.00*
0-521-52415-6 Paperback $23.00*

Utopia
Second Edition
Thomas More
George M. Logan and Robert M. Adams, Editors
Cambridge Texts in the History of Political Thought
0-521-81925-3 Hardback $30.00
0-521-52540-3 Paperback $10.00

Upheavals of Thought
The Intelligence of Emotions
Martha C. Nussbaum
0-521-53182-9 Paperback $27.00*



MARTHA C.
NUSSBAUM
UPHEAVALS
OF THOUGHT

Visions of Politics
Vols. 1-3
Quentin Skinner
0-521-81382-4 Hardback $180.00
0-521-89075-6 Paperback $65.00

The Philosophy of Social Practices
A Collective Acceptance View
Raimo Tuomela
0-521-81860-5 Hardback $60.00

Vico: The First New Science
Giambattista Vico
Leon Pompa, Editor
Cambridge Texts in the History of Political Thought
0-521-38290-4 Hardback $65.00
0-521-38726-4 Paperback $23.00

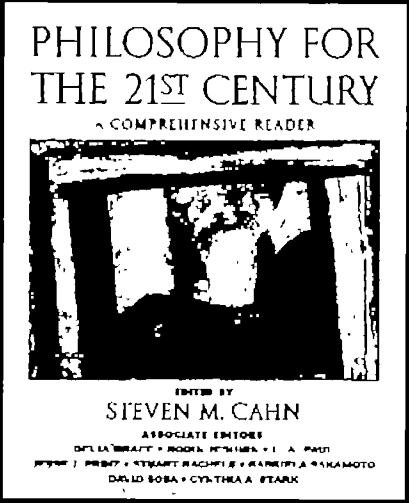
Self Love and Christian Ethics
Darlene Fozard Weaver
New Studies in Christian Ethics 23
0-521-81781-1 Hardback $65.00
0-521-52097-5 Paperback $23.00

The Cambridge Companion to Duns Scotus
Thomas Williams, Editor
Cambridge Companions to Philosophy
0-521-63205-6 Hardback $65.00
0-521-63563-2 Paperback $23.00

*prices subject to change

Available in bookstores or from
CAMBRIDGE
UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

New from the
OXFORD HIGHER EDUCATION GROUP



PHILOSOPHY FOR THE 21ST CENTURY
A COMPREHENSIVE READER
EDITED BY
STEVEN M. CAHN
ASSOCIATE EDITORS

[illegible] and philosophy of art.

This volume is unique in drawing on the judgments of a new generation of scholars, each of whom has chosen the articles and provided the introduction for one section of the book. These associate editors—Delia Graff, Robin Jeshion, L. A. Paul, Jesse J. Prinz, Stuart Rachels, Gabriela Sakamoto, David Sosa, and Cynthia A. Stark—are at the forefront of 21st-century philosophy. Their selections include the work of such leading contemporary thinkers as Nancy Cartwright, Saul A. Kripke, David Lewis, Thomas Nagel, Robert Nozick, Derek Parfit, and Sydney Shoemaker, along with classic works from 2500 years of philosophy. The book has been structured to maximize continuity, and an introductory essay by Simon Blackburn explains the tools of symbolic logic. This groundbreaking volume sets a new standard for introducing students to the importance and fascination of philosophical inquiry.

2002 · 864 pp., 3 illus. · paper

OXFORD [illegible]
1-800-451-7556
[illegible]

THE JOURNAL OF
PHILOSOPHY

19/5/03

VOLUME C, NUMBER 3
MARCH 2003

page
111 *Qualia and Analytical Conditionals* David Braddon-Mitchell
136 *The Scope Argument* Michael O'Rourke

COMMENTS AND CRITICISM
158 *Savoir Faire* Ian Rumfitt

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER. Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2003 by the Journal of Philosophy, Inc.

ISSN 0022-362X

9/5/03

THE JOURNAL OF PHILOSOPHY

VOLUME C, NO. 3, MARCH 2003

QUALIA AND ANALYTICAL CONDITIONALS*

By the end of the last century, it started to look like the philosophy of mind had stark choices to make. Accept the idea that intuitions give some kind of guide to possibility, and dualism seemed to be just around the corner, or so David Chalmers[1] told us. Deny that intuition gives any guide to possibility, and we can dismiss easily Chalmers's "Zombie" argument and its cousins[2]—but then, as Frank Jackson[3] has taught us, we have no explanation, paradoxically, for the core arguments for the relevance of the a posteriori in metaphysics. The best explanation depends crucially on thought experiments—intuitions about possible cases—like twin earth.

Here is a reminder of the story so far. Intuition seems to tell us that "Zombies" are possible: physical duplicates that lack qualia.[4] Assume we have qualia. If there is a possible world that duplicates this one physically and only physically, and that lacks qualia (which is

---

*This paper has benefitted enormously from input from David Chalmers, Frank Jackson, Fred Kroon, Kristie Miller, Philip Pettit, Denis Robinson, and Laura Schroeter

[1] *The Conscious Mind* (New York: Oxford, 1996).

[2] Chalmers is the source of the current popularity of the Zombie arguments. Keith Campbell also has a nice version of the argument in *Body and Mind* (New York: Doubleday, 1970). The term seems to enter philosophy via R. Kirk, "Zombies versus Materialists," *Proceedings of the Aristotelian Society*, XLVIII (supplement) (1974): 135–52.

[3] *From Metaphysics to Ethics: A Defense of Conceptual Analysis* (New York: Oxford, 1998).

[4] Throughout this paper I am using 'qualia' in the sense that is neutral as to whether they are physical states or not. Of course 'qualia' is a technical term, so the sense in which I am using it has, if we are to find a folk analysis, to be thought of as a synonym for nontechnical notions like 'sensory experience', or 'raw feel' stripped of any particular philosophical theory about such experiences. There is, of course, another terminological tradition in which the name is reserved for nonphysical states, and the debate is whether there are any.

0022-362X/03/0003/111–35

what a Zombie world amounts to) then we have a minimal physical duplicate[5] of actuality that fails to be a duplicate *simpliciter*, and so physicalism is false. The complete way things are physically would turn out not to be sufficient for qualia.

How should the physicalist respond? One way is to rule out such a case on a priori grounds. What I might call *procrustean analytical functionalism*[6] does this: it is just a priori on proper reflection, that mental states are what play the relevant causal functional roles, and so a thorough understanding of the meaning of mental language rules out the logical possibility of the roles being played in the absence of mental states. Aside from its implausibility, this kind of functionalism embraces the a priori presumably in part through reflection on intuitions about cases, and thus would need to give an account of why the fairly robust intuitions about the possibility of Zombies is trumped by the apparently more recherché intuitions which ground its claims for analyticity. It seems arbitrarily to privilege certain intuitions.

Probably the most prominent response is to deny that intuition is any guide to possibility at all.[7] After all, the a posteriori physicalist tells us, it may have seemed intuitive that it was impossible for water to be black and tarry. But we have learned from Saul Kripke, Hilary Putnam, and others[8] that this is false, since if H_2O were to be black and tarry it would be water nonetheless. Equally so, they would argue, for physicalism about the mind. Neuroscience will tell us what minds are, and thus tell us what is necessary. Any intuitions about the possibility of mental states not of this appropriate kind are in the same prescientific class as those we might have had about the possibility of non-H_2O "water" before the nature of water was well understood. So there are necessary truths, such as the nature of water, which since they are a posteriori do not depend on intuitions. So intuition is not a guide to possibility or necessity.

---

[5] See David Braddon-Mitchell and Jackson, *The Philosophy of Mind and Cognition* (Cambridge: Blackwell, 1996).

[6] I suppose this includes David Lewis, "Self-Profile," in Samuel Guttenplan, ed., *A Companion to the Philosophy of Mind* (Cambridge: Blackwell, 1994), pp. 412–31; Sydney Shoemaker, "Functionalism and Qualia," *Philosophical Studies*, xxvii, 5 (1975): 292–315; and Braddon-Mitchell and Jackson. I am supposing that everyone who is an analytical functionalist is also a procrustean analytical functionalist.

[7] Michael Tye, "The Subjective Qualities of Experience," *Mind*, xcv (1986): 1–17, see pp. 5–6; Stephen Yablo, "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research*, liii, 1 (1993): 1–42; and Robert Stalnaker and Ned Block, "Conceptual Analysis, Dualism, and the Explanatory Gap," *Philosophical Review*, cviii (1999): 1–46.

[8] Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980); Putnam, "The Meaning of 'Meaning'," in *Selected Papers*, Volume 2 (New York· Cambridge, 1975), pp. 215–71.

But this is too simple. There is a compelling account of how the necessary a posteriori works according to which in each case the arguments for necessity depend on an intuition about possibility.[9] In the case of water, an intuition that what on twin earth failed to share the chemical structure of the watery stuff we are in contact with would not be water, in the case of mental states the intuition that what failed to share neuro-functional properties with paradigm minds would not count as mental. So this approach could be redescribed as in fact depending on an analytic claim. In the case of qualia, it would be the claim that qualia are whatever neural state types are in fact actually causally responsible for our discriminative powers. Once we have added this additional analytic premise, we can see that it is very similar to analytical functionalism: it just has an actually clause added to make it rigidify on the actual occupants of the role.[10] The claim that intuition is no guide to possibility turns out to depend on intuitions about possibility, and thus is either inconsistent or else open to the charge of choosing between intuitions arbitrarily.[11]If we accept *those* intuitions as being a guide to the meaning of 'qualia', why reject the Zombie intuition?[12]

Out of all this, there emerge a number of claims that seem right to me. First, physicalism is, I think, probably true, and some kind of substantive functionalism is my preferred brand. Mental states are, I think, whatever play certain causal roles (or if you prefer, whatever is intrinsically similar to that which actually plays the roles: for current purposes this is an in-house dispute). So the first claim will be:

(1) Substantive functionalism about qualia: qualia are whatever play certain roles.

But, I do not intend this claim to be directly analytic. So that gives us (2).

(2) The claim that 'qualia are whatever plays certain functional roles' is not directly analytic.

---

[9] See Jackson, and, for a contrary view, Robert Stalnaker, "On Considering a Possible World as Actual," *Philosophical Studies* (forthcoming).

[10] L. Humberstone and M. Davies, "Two Notions of Necessity," *Philosophical Studies*, xxxviii (1980): 1–30.

[11] This is not uncontroversial: see Stalnaker.

[12] Of course, one could tell a story about equlibration, in which the Zombie intuition is reluctantly cast aside in favor of a larger consistent set, but that underplays the vividness of it for many

On the balance of probabilities, physicalism is true, so:

(3) Whatever plays those roles is physical, so the qualia around here are physical states.

I agree with Chalmers that the Zombie intuition is powerful, and that modal intuitions play a role in settling meaning.

(4) The intuition about possibility that underlies Chalmers's Zombie arguments has as good a claim to be taken as a prima facie guide to possibility as any other equally robust intuition. No story that tells us that Zombies are impossible can do so without explaining why this intuition is misleading.

(5) In general, intuitions (or perhaps the dispositions that the intuitions are a fallible guide to) are a guide to possibility via analyticity.

These five principles seem difficult to hold together. Agreeing that we have the Zombie intuition and that it is a guide to possibility seems to rule out (3). And (2), which says that it is not directly analytic that functionalism is true, undermines the usual reasons for holding (1).

I think these claims can in fact be reconciled. I plan initially to explain the reconciliation by raising the problem in a parallel case: the example of water that has occupied so much of the literature.

I. WATER AND AQUEOUS ESSENCE

Suppose that things were a little different from how they are with respect to the word 'water'. Imagine we speak a slightly different language (and have slightly different thoughts) in which the following claims are true. Although H_2O is in fact water, once it was thought that there was an "aqueous essence"—a nonphysical property distinct from functional properties like fluidity, clarity, potability, and so forth—which was what made water the substance that it is. It was not that it was thought that aqueous essence was the nonphysical cause of clarity, fluidity, and so on. These properties were thought to be *markers* of the presence of aqueous essence if there was any. Aqueous essence was thought to be causally idle, much as parallelists about the mind think the mind is causally idle. And suppose that this thought was still present enough in people's thinking about water that they were prone to have intuitions that something could be a physical duplicate of water without being real water (in virtue of lacking the aqueous essence). I do not think the concept of water is or was like this, but my purpose here is to show how such a concept would be both coherent, and compatible with physicalism about water, understood in this different way.

To show that this variant concept is both coherent and compatible, we will have to provide some sort of analysis of 'water'. If it is analytic that water has to possess some nonphysical aqueous essence, then, of course, if physicalism were true there would be no water. So, any such analysis would not have the coherence I am suggesting. Suppose, on the other hand, the right understanding of 'water' was what may actually be the correct understanding: the chemical properties that explain the functional properties of the paradigm samples of water. Then if physicalism were true, the intuition about possibility would be just mistaken, and thus the package I am trying to make consistent equally would shatter.

But what about the following analysis?

> AE: If there is actual aqueous essence, then that which possesses it is water and necessarily so. Otherwise, water is whatever shares explanatory chemical features with paradigm samples.

This is what I shall call a *conditional analysis*. There is a sense in which the right substantial analysis depends on how things actually are. But grasp of the concept does not depend on knowing how things actually are, hence the concept could be grasped before knowing either whether there is any actual aqueous essence or even what the underlying chemical properties of the paradigms are.

How should we understand a conditional analysis like this? I propose to use the strategy of *two-dimensionalism*:[13] using a two-dimensional matrix, where we consider different hypotheses about how the actual world is,[14] and then consider what sets of counterfactual judgments about logical possibility we would make on various presuppositions about the actual world.

Usually, of course, the different ways actuality is conceived to be are variant physical hypotheses (the existence of XYZ—understood as physical but perhaps counternomic, and so on). But there is nothing to prevent us from considering ways things could be considered as actual that include the presence of strange possibilities such as nonphysical aqueous essence. After all, none of us should assign a *zero* credence to the claim that things are in fact this way!

I.1. Evaluation of 'water is H_2O'. So let us evaluate some claims about water under these varying circumstances. Let us start with 'water is

---

[13] Stalnaker, "Assertion," in P. Cole, ed., *Syntax and Semantics: Pragmatics*, Volume IX (New York Academic, 1978); Humberstone; Chalmers, Jackson.

[14] I write here in the indicative mood, which suggests an understanding of two-dimensionalism according to which one dimension is purely epistemic possibility, and the other metaphysical (Chalmers prefers such an analysis in his forthcoming "The Foundations of Two Dimensional Semantics"). Nothing here depends on this reading.

H_2O'. I shall illustrate this evaluation with a matrix consisting of rows each of which illustrates a different hypothesis about actuality (the hypothesis on the row headings). Each box in the row illustrates what to say about the expression being evaluated in a counterfactual world, the relevant features of which are described in the column heading. I shall not provide further matrices for the other expressions; this example will, I hope, be enough to help those who are in general helped by diagrams.

The headings on rows and columns 'AE' mean that there is some aqueous essence in the world (that is, the nonphysical substance copresent with whatever is responsible for clarity, potability, and so forth). The headings 'H_2O' and 'XYZ' stand for those physical substances that are what is potable, clear and so forth—in short, the substances that play the water role. By hypothesis about the deviant concept, if there is any AE, water is AE, not that which plays the water role.

	AE, H_2O	AE, XYZ	−AE, H_2O	−AE, XYZ
AE, H_2O	F	F	F	F
AE, XYZ	F	F	F	F
−AE, H_2O	T	T	T	T
−AE, XYZ	F	F	F	F

'Water is H_2O'

In all the worlds considered as actual in which aqueous essence is present, water is not H_2O. From the perspective of *these* worlds, this is a necessary claim: all and only the counterfactual worlds that contain AE contain water, and all and only the AE across all worlds is water. So all the rows in which there is actual AE are ones where 'water is H_2O' is false in every box: it has a necessarily false *C-intension* at these rows.

In general, a *C*-intension of a term is a function that assigns to each world the extension of that term on the assumption that we keep the actual world fixed as whichever one is in fact actual—it is close to the ordinary use of 'intension'. Two-dimensional diagrams illustrate the point that since this function can vary according to context, there are as many *C*-intensions as there are contexts that we can consider to be actual. The expression *C*-intension (for counterfactual intension) is due to Jackson (*op. cit.*, pp. 48–49), and is motivated by the idea that most of the worlds in a *C*-intension are thought of as counterfactual:

the first world in a row of a table is thought to be actual, and the other worlds are from its perspective mere counterfactual possibilities. Some may be familiar with the idea from Chalmers's expression *secondary intension*.

There are some ways things could be *actually* where 'water is H_2O' is false, as well as some ways things could be *actually* where it is true. So, in the language of two-dimensionalism, we say that 'water is H_2O' has a contingent *A-intension*.[15] The diagonal that we can draw where we consider only the worlds considered as actual and evaluate the expression in those worlds contains boxes where we assign 'T' to 'water is H_2O' and boxes where we assign 'F'. This is in virtue of the first clause of the analysis. We were entitled to ignore the (physical) intrinsic nature of whatever played the water role,[16] because if there is AE then AE is water and necessarily so.

In general, an A-intension of some term is the function that assigns to each world the extension of that term *on the assumption that that world is itself actual*. Chalmers uses the expression *primary intension* for this notion.

So let us consider the worlds (considered as actual) where there is no AE. In these worlds, the conditional analysis directs us to consider what physical properties play the water role. If a world considered as actual has no AE and it has H_2O playing the water role, then H_2O is water *and necessarily so*. From the perspective of a world considered as actual without AE, any counterfactual worlds that contain no H_2O and *do* contain some mysterious nonphysical AE property do *not* contain water, they merely contain a weird alien property. Thus, if things are this way actually, then we should judge it to be necessary that water is H_2O. Similarly, if actually the water role is played by XYZ, then from that perspective the claim 'water is necessarily XYZ' is true.

What is the take-home message? That 'water is H_2O' comes out true and necessarily so on the supposition that actually H_2O plays the water role and there is no AE, and false otherwise. In the language of two-dimensionalism, on this supposition, 'water is H_2O' has a necessary C-intension.[17]

I.2. Evaluation of 'water is the watery stuff'. Consider now the claim that water is the watery stuff (that is, what plays the water role is water). This comes out as false when we consider worlds considered as actual that contain AE—for AE plays no causal role. And it comes out as necessarily false from the perspective of these strange worlds, for worlds considered as counterfactual are judged by the same standard as the world taken to be actual: all and only the causally idle AE is water.

---

[15] For Chalmers, *contingent 1-intension* or *primary intension*; for Gareth Evans and Humberstone and Davies, *deeply contingent*.

[16] That is, fills the lakes, the river, taps, and so on.

[17] The terminology gets baroque: for Chalmers, it is a *necessary 2-intension* or *secondary intension* on this supposition. On some understandings, this corresponds to what Humberstone and Davies might state by "'water is H_2O' is superficially necessary."

What, though, of the worlds in which naturalism is true? Well, if H_2O plays the water role actually, then all and only the worlds where H_2O plays the water role are worlds where water plays the water role (a world considered as counterfactual from this perspective where XYZ plays the water role is a world where the water role is played by something other than water). But from the perspective of worlds considered as actual where XYZ plays the water role, all the worlds where H_2O plays the water role are worlds where something other than water plays the water role. So 'water plays the water role' is not necessary from any perspective, for there is always a world considered as counterfactual where it is false.

Sometimes at this point in a discussion of two-dimensionalism about water that does not consider nonnatural worlds considered as actual, the claim is made that, although never necessary, we can see from the analysis that 'water is the watery stuff' is true a priori. This is because, however the actual world is, if something plays the water role, the analysis says that it is water. So we know that—surprise, surprise—water plays the water role (if there is any). This corresponds to 'water plays the water role' being true down the diagonal of a two-dimensional diagram.[18] This will not be true on our conditional analysis, for in the nonnatural worlds that contain AE, water is not what plays the water role, even if H_2O does. But there is a strange kind of conditional thing that we can know a priori on this model: if there is no AE actually, then water plays the water role. The apparent a priority of 'water plays the water role' does, however, correspond to its being true in all the naturalistic worlds where the same world is treated as world of index and of context. Thus, for this imaginary variant concept of water I am using to illustrate the idea, 'water plays the water role' has a contingent *A*-intension, though it has a restrictedly necessary *A*-intension whenever the actual world has no aqueous essence.

I.3. Evaluation of 'water is AE or it is whatever plays the water role'. Lastly, consider 'water is AE or it is whatever plays the water role'. This does come out as having a necessary *A*-intension (and thus is plausibly a priori):[19] there is no world considered as actual in which we could deny it, even though it is not necessary.[20] Consider the world as actual in which H_2O plays the water role. In that world, water plays the water role, so the disjunction is true. But consider a counterfactual world with no AE and in which XYZ plays the water role: neither disjunct is true. So the disjunction is not necessary, even though a priori. The disjunction is the equivalent of the contingent a priori claim 'water plays the water role' on a more standard understanding of 'water'. On this toy analysis, 'water is AE or it is whatever plays the water role' has a necessary *A*-intension.

---

[18] And thus has a necessary *A*-intension or 1-intension.

What all this shows is that a concept of water could coherently be one which prioritized the nonphysical, consistently with physicalism about water being both true and necessarily true. Nonetheless the concept's nature explains an intuition that there could be something physically like water but which fails to be water. This is because we ought to grant some credence to the thought that the actual world contains aqueous essence; and if it does, then the very same *analysis* of water would tell us that there can be physical duplicates of water that fail to be water. Of course, given how we think things actually are—where we place most of our credence—the analysis tells us to use 'water' in a way that makes it impossible for a physical duplicate to fail to be water. But important though intuitions are in buttressing arguments for possibilities, we should not expect them to do anything quite so fine-grained as distinguishing between a genuine possibility that a physical duplicate of water could fail to be water, and a credence in things being actually such that, with the same *concept* of water, the *content*[21] of claims about water is different enough to allow that physical nature does not settle what counts as water.

I propose to do something similar in the case of qualia. I shall develop a conditional analysis of 'quale' which explains the intuition that Zombies are possible, by showing how the very same *concept* of quale gives rise to a different content should physicalism not be in fact true in the actual world.

---

[19] See Chalmers, "The Foundations of Two Dimensional Semantics."

[20] I use 'necessary' to correspond to the necessary *C*-intension under some supposition about actuality, what Evans called "superficial necessity." This terminology confuses some these days, many naturalists are inclined to think anything a priori is superficial, and necessities like 'water is H_2O' are deep

[21] Terminological footnotes must abound in discussions of two-dimensionalism. I use 'content' in roughly David Kaplan's sense—see his "Dthat," in P. Cole, ed., *Syntax and Semantics* (New York: Academic Press, 1979), pp. 221–53, and his "Demonstratives," in Joseph Almog, John Perry, and Howard Wettstein, eds., *Themes from Kaplan* (New York: Oxford, 1989) pp. 481–563; equivalent perhaps to Jackson's *C*-intension, and Chalmers's secondary intension.

II. A CONDITIONAL ANALYSIS OF 'QUALE'

An important mistake, which has bedeviled the philosophy of mind, is to try to use conceptual matters to determine whether physicalism is true or not. I take that to be an empirical matter or else one of substantial metaphysics.[22] I do not plan to directly argue for this claim here, instead I shall recommend a procedure that reverses the direction of inquiry, and hope the benefits this reversal brings are argument enough.

Instead of asking how our concept settles the truth or otherwise of physicalism, we should ask what our concept tells us about the mind, should there actually be nonphysical states (say, ones whose intrinsic nature is grasped in experience), and what it tells us about the mind should physicalism be true.

Suppose then that, contra my expectations, it turns out that there are special properties left out of the physical picture of the world, whose intrinsic nature (and perhaps no other intrinsic natures[23]) are grasped in experience. I shall call such states henceforth the *spooky states*. If there are spooky states—and they are the things whose nature I grasp when I experience a quale of red—then those things are indeed the qualia of redness. If there were such states, I think we would be inclined to take it that they are necessary for having qualia. That this is somehow conceptually important for the nature of qualia is evidenced in the tendency for people to default to a kind of dualism in the absence of arguments for physicalism.

But what if it turns out that there are no spooky states? That is pretty much the consensus in the natural sciences, and yet this alone does not seem to lead to the abandonment of the idea that we have qualitative experience; it rather leads to attempts to account for qualitative experience physically. I think there is consensus among many materialist philosophers of mind that the states which play complex functional roles which underlie our discriminatory capacities and sensory talk are in fact our qualia. There is less agreement on what it takes to have qualia for beings unlike us. Analytical functionalists (and those who hold successor views like mine) think that what it takes is to have those roles occupied, and that in each case the qualia are the occupiers of the roles.[24] Others—who hold the view that Jackson and I have called *empirical functionalism*[25]—think that what it takes is to have states similar in some way to the states that occupy the roles in us (the debate being as to the level and degree of similarity: physical, neural, psychofunctional, computational, and so on). But what if, counterfactually, we consider worlds that have spooky states? On the assumption that actually there are none of these spooky states, strange nonphysical states in other possible worlds do not get to count as qualia unless they play the functional roles. For strange nonphysical states that fail to play the functional role would count as what I mean by 'quale' only if they are what I am *actually* directly acquainted with in experience.

---

[22] This does not rule out the possibility that some versions of dualism, if false, are a priori false—because there is some hidden incoherence. But even in this case what I can consider can be reconstructed with worlds considered as actually understood as a kind of epistemic possibility—in the same way as there is *some* sense in which it is epistemically possible that Fermat's Last Theorem is not a theorem (maybe there is a flaw in the proof).

[23] Lewis, "Ramseyan Humility" (forthcoming).

So here, then, is an analysis:

If there are spooky states then

- In the actual world the qualia are the spooky states; and all and only the qualia in counterfactual worlds are the spooky states

Else

- In the actual world the qualia are the states that play the functional roles; in other worlds, qualia (if any) are the states that play the roles *in that world.*

Anyone who is a friend of empirical functionalism should feel free to substitute the following for the second bullet point: the issue of the rightness of that analysis has been argued elsewhere (*ibid.*) and, again, nothing in this paper depends on it.

- In the actual world the qualia are the states that play the functional roles; in other worlds the qualia, if any, are the states that play the roles in the actual world.

III. SOME DEFENSE AND SOME CLARIFICATION

The analysis we have provided so far uses possible worlds *simpliciter.* It does not refer to locations in a world. Unfortunately, this is too much of a simplification to be of use except as preliminary exposition. We have not said where in a world the nonphysical states occur in order to count as satisfying the antecedent.

In addition, once we specify this, we will need both to consider worlds considered as counterfactual with the spooky states located at various points, and to consider worlds considered as actual with spooky states located at various points.

---

[24] Though some say that they are the second-order state of the role's being occupied.

[25] Braddon-Mitchell and Jackson, "The Divide and Conquer Path to Analytic Functionalism," *Philosophical Topics*, XII, 1–2 (1999): 71–88.

Let us begin with the counterfactual claim. What if it were the case, contrary to fact, that George Bush were the only person in the world who had the appropriate spooky states? Let us start by evaluating this claim on the assumption that in the actual world there are no spooky states anywhere. In this case, I claim that the right thing to say is that in the counterfactual world, there is a strange nonphysical property that is not the property of being a quale. It is just a world where Bush is stranger than he is actually. But conceived of as *actual*, the world where Bush has spooky states is one where he has qualia. So the special conceptual priority of the nonphysical properties is not one that extends to counterfactual worlds. It is only if *actually* there are spooky properties as well as functional or physical ones that the concept of quale latches onto the spooky ones.

But what to say of a world conceived of as actual where Bush has the spooky states but I do not? It would be disturbing if this counted as a case where I discover that I lack qualia. This would mean that in principle I could make the discovery that I lacked qualia (and thus perhaps could legitimately be tortured) by discovering nothing about myself, but rather by discovering strange metaphysical truths about Bush.

On the other hand, the analysis needs to give the answer that a world conceived of as actual where I possess the spooky property is one where it is true that qualia are spooky, and necessarily so (in the sense of necessary C-intension).

How to pull off this trick? One possibility is to do it all with averages. 'There are spooky states' means something like "it is usual in the paradigm community for there to be properties that obey the *identification hypothesis*" or something like that. This however leaves one open to various kinds of implausible contextualism.[26] In particular, if actually most people have the property, and I do not, then I discover I lack qualia.

The right solution is to accept that the concept of qualia has a hidden indexical structure. My use of the term is governed by a presupposition that I cannot fail to have qualia if it seems to me that I do, even though I could counterfactually lack qualia. (Relative to an actual world in which I have the spooky property I could lack qualia by lacking the spooky property. Relative to actual worlds where there is no spooky property I could lack qualia by being, say, a Peacockian Martian Marionette.[17])

---

[26] Cf. some of the issues in Lewis, "Mad Pain and Martian Pain," in his *Philosophical Papers*, Volume I (New York: Oxford, 1983), pp 122–33.

So, in my mouth the term 'quale' functions thus: if there is a nonphysical property whose intrinsic nature I am acquainted with in experience, then that is my quale, and it is a necessary condition for anything to be a quale that it be a property like that. If not, then my quale is whatever state plays a certain causal functional role, and this is sufficient for anyone to have a quale.

The egocentricity captures, I think, what is acceptable about the problem of other minds. I can think of someone else that he probably does have qualia. But on the somewhat mysterious assumption about the nature of qualia that would be right in a world conceived of as actual in which there are spooky states, it is an open question for me. Of course, assuming physicalism is actually true, the functional analysis deflates the problem of other minds, for I have reasonable epistemic access to others being in states which play the roles in them. Therefore, whether there is a problem of other minds will depend on whether (actually) there is a spooky state with which I am acquainted. So the worlds conceived of as actual need to be *centered worlds*, and the claim that 'there are spooky states' needs to be evaluated at their centers.

There is still one final comment to make on these matters: of course if the way things actually are is one in which Bush and no-one else has the spooky states, then he, Bush, can claim and claim truly that the rest of us are all Zombies. But of course the content of 'Zombie' in his mouth is a little different from the content of Zombie in mine, when I claim in my idiolect that I am not a Zombie (even though both our terms function similarly). This is something of a cost, but I think that it does reflect what remains of the significance of the problem of other minds. The problem arises exactly because there is something idiolectic about all of our uses of terms like 'quale'. And the problem fades on plausible assumptions about how things are that ensure that the idiolectic differences come to nothing.

We can now modify the analysis with which we began by relativizing in the following way:

For some speaker, S (who is the center of a centered world):

If there are spooky states at *the center of the actual world* then

---

[17] Christopher Peacocke, *Sense and Content* (New York: Oxford, 1983), see pp. 204–05.

- In the actual world all qualia are the spooky states; and all and only the qualia in counterfactual worlds are the spooky states.

Else

- In the actual world the qualia are the states that play the functional roles; in other worlds qualia (if any) are the states that play the roles in that world.

IV. DOING SOME TWO-DIMENSIONAL MATRICES WITHOUT TABLES

Analogously with the case of water, I will now evaluate a few propositions associated with qualia to see how they behave in two-dimensional contexts.

IV.1. Evaluation of 'qualia are NS1 states'. I shall start with the evaluation of a physicalist claim about the nature of qualitative states. Suppose that some physical state or states are found to underlie the typical functional roles (in recognition and manipulation and so on) of qualia. Call these the *physical correlates* of qualia. On some physicalist doctrines, these count as necessary conditions for qualia.[28] Let us use the label 'NS1' for some neural state that correlates with qualia.[29] On these sorts of physicalisms 'qualia are NS1 states' would be true necessary a posteriori. This empirical functionalist reading is not something with which I agree; however, nothing here hangs on this as I said above.

Under what conditions is it true that qualia are NS1 states given our analysis? If in the actual world there are relevantly placed spooky states, then qualia are never NS1 states. So in a two-dimensional matrix 'qualia are NS1 states' has a necessarily false C-intension in all the rows where actually there are spooky states. In the worlds considered as actual where there are no spooky states, it is irrelevant whether or not qualia are NS1 states in the actual world: qualia are NS1 states in every world where NS1 states play the functional role regardless of what play the roles actually.[30] So 'qualia are NS1 states' is true down a *part* of a column in which we look at worlds where qualia play the NS1 role: that part of the column in which the designated or actual worlds are non-spooky.

Still, there is no row in which it is always true, and nor is it true down the so-called diagonal: there are worlds considered as actual in which it is false (any world where either NS1 states do not play the role, or in any world at all when the actual world has spooky states). Thus this claim is what Gareth Evans[31] called *deeply contingent*—indeed it is contingent simpliciter.

---

[28] Braddon-Mitchell and Jackson, "The Divide and Conquer Path to Analytic Functionalism."

[29] Of course this is an absurd simplifying assumption: every quale type will correlate with a different physical type, and, for all we know, every token as well.

[30] This is the contrast with empirical functionalism, where NS1 states are qualia if and only if they actually are.

IV.2. Evaluation of 'qualia are whatever play the functional roles'. For what I have called procrustean analytical functionalism, the claim that 'qualia are whatever play the functional role' is analytic. It is true right down the diagonal—in every world considered as actual. It is also necessary. Even empirical functionalism gives us an *A*-intension in which it is necessary (though then it will not have necessary *C*-intensions because in every row there will be worlds where what plays the role is not what plays the role in the actual world). On the conditional analysis we are considering, however, 'qualia are whatever play the functional roles' will come out as having neither a necessary *A*-intension nor in general necessary *C*-intensions. For if in the actual world there are spooky states, then at any world where there are spooky states that do not play the causal functional roles, the states playing the roles are not qualia. Thus, the claim that 'qualia are whatever play the functional roles' has a contingent *A*-intension. On the (controversial) assumption that necessary *A*-intension amounts to analyticity, we can see that on this assumption the constitutive claim of procrustean analytical functionalism comes out false. It is not analytic that qualia are whatever play the functional role.

And yet the conditional guarantees something like analytical functionalism: the part of a two-dimensional matrix that leaves out all the rows where the actual world has spooky stuff is the same as the matrix for analytical functionalism. On the assumption that the actual world is physicalistic, we are required to regard whatever plays the role (actually and counterfactually) as qualia. But what kind of claim is this? Remember that I have rejected bare necessities. There needs to be some analytic claim that grounds everything. There needs to be some proposition with a necessary *A*-intension that will do the job for my conditional analytical functionalism that the claim that 'qualia are whatever it is that plays the role' does for procrustean analytical functionalism.

IV.3. Evaluation of 'there are spooky states and they (and only they) are the qualia, or there are no spooky states and the qualia are the states that play the functional role'. This is the claim that does the job. No matter how things are actually, I know that I am either in direct contact with the essential properties of spooky stuff in my phenomenal experience, or I am not. If I am, I know that they are the qualia (so the conditional that if there are spooky states, then 'they are qualia' is true at every world down the diagonal) and I know that if there are no spooky states, then the states that play the role are qualia (so the claim that if there are no spooky states then qualia are the states that play the role is true down the diagonal). All of this is part of a priori analysis. I hope that this article will do part of that job—the part of persuading the reader that if there are such spooky states, then they are qualia, and elsewhere I have a priori arguments for analytical functionalism that can be modified to support the more limited claim that if there are no spooky states then qualia are the states that play the role.[32]

---

[31] "Reference and Contingency," *The Monist*, LXII (1977): 161–89.

V. WHY THIS EXPLAINS THE ZOMBIE INTUITION

The point of this somewhat complicated analytical framework is not just that it is independently correct. For our current purposes, the benefit is that it gives an account of our concept of qualia that allows for the truth of physicalism while explaining the Zombie intuition: the intuition that there could be a physical duplicate that lacked qualia.[33]

Why is this so? Well, technically the answer is clear. The claim that 'there is a physical duplicate of us with no qualia' has a contingent *A*-intension on this analysis. For there is a world considered as actual where the spooky property exists (say I am an ectoplasmic being with no causal connection to the world, and am directly in contact with the intrinsic nature of experience). In this world, if there are any physical beings floating around that answer to the physical descriptions we give of ourselves,[34] and are unconnected to the spooky properties, they lack qualia. And there are ways things could be actually in which although we are physical beings and possess qualia (because of psychophysical laws that connect the physical to the nonphysical) it would be true, if things are actually that way, that this would be logically contingent, since if that world were actual then there would be counterfactual worlds in which the laws are missing. It would be true to say that if we were not connected to the nonphysical in that way, we would lack qualia. So there is a way things could actually be, according to which it is true that I do not lack qualia, but it is possible that my physical duplicate would (say, in worlds where any law-like connection between the physical and the nonphysical is broken).

---

[32] Braddon-Mitchell and Jackson, "The Divide and Conquer Path to Analytic Functionalism"; and Braddon-Mitchell and Jackson, *The Philosophy of Mind and Cognition.*

[33] Strictly speaking, this is an absent qualia case. Chalmers thinks of Zombies more generally as lacking consciousness: more on this later.

[34] There is a complexity here: epistemically, if I am actually ectoplasmic, what do we say about the physical? Plausibly on an indicative reading of the conditional, the physical should be as it seems to be, and the ectoplasmic difference as small as possible reconciled with my knowledge.

The physicalist says, however, that things are not actually this way. The way things are is a way that makes it false to say that a physical duplicate of mine could lack qualia. On physicalist assumptions about actuality, 'Zombies cannot exist' has a necessary *C*-intension. The contingent *A*-intension explains the intuition that it is possible; the necessary *C*-intension explains the sense in which it is not.

Another way of putting this has to do with the fine-grainedness of modal intuitions. While I think they are important, it is not clear that they can distinguish two cases. An intuition is a kind of picture that makes sense: I picture a world with physical duplicates that lack qualia, and it does not seem to be incoherent. But of what is it a picture? I do not think that intuition gets this clear. Is it a picture of what is possible, given how things are; or does it correspond to the fact that as a physicalist, but hopefully not an arrogant one, I do not give zero credence to physicalism being false? And if I do not give a zero credence to this, I cannot give a zero credence to the expression in my mouth 'Zombies are possible' expressing a truth—even though it does not do so given how things are, and cannot do so. Intuition does not distinguish between a non-zero credence in things being such that 'there could be physical duplicates that are Zombies' might be rightly thought true, and a direct intuition that it does express a truth given how things actually are.

This begins to explain why there is such a dispute in the philosophy of mind about whether Zombies are possible. I think it also explains more fine-grained features of the dispute. It is no accident that it tends to be physicalists who lack the intuition that Zombies are possible, and dualists who have it. The uncharitable explanation is that each has intuitions designed just to shore up their view. But a two-dimensional understanding of the concept gives a better explanation. A committed physicalist has a firm view about what world is actual. For her, the relevant possibilities are about what is possible given how things are, and if things are the way physicalism says, then Zombies are impossible (given the necessary *C*-intension). The committed dualist thinks there are actual spooky properties, and that they are our qualia. She thinks she lives in a world where 'Zombies exist' has a contingent *C*-intension. So, the possibility of Zombies is something to be defended strongly. Thus, on the second understanding of the concept, we should expect anyone with firm intuitions about how things actually are—about whether physicalism is actually true or false—to have strong intuitions in different ways about the possibility of Zombies: for they differ in the *C*-intension that they assign to the sentence. And, of course, someone who is undecided about the truth of physicalism will tend also to defend the claim that Zombies are possible, but for a different reason. If someone is undecided about the truth of physicalism, then she entertains alternative hypotheses about how things actually are, and thus will tend to concentrate on the *A*-intension. Since the *A*-intension is contingent, it will seem as though 'Zombies are possible' is true *simpliciter*, and so without the aid of the clarification provided here, it will seem like an argument for dualism. The upshot is: some intuitions about what is actual are driving some intuitions about what is possible. Thus it is a kind of circularity to argue from those intuitions about what is possible to claims about what is actual.

V.1. An objection: "this just says we might have meant something else." Here is an objection that has been pressed on me by many. It runs thus: what you have explained is not why we have an intuition that Zombies are possible. You have only shown that one might be a physicalist and think that we might have meant something else by Zombie such that it would have been right to say that Zombies are possible. And, of course, we always knew that in the world where 'Zombie' means "Zebra," it would be right to say 'Zombies are possible'. And epistemically 'Zombie' might (though it is very unlikely) mean "Zebra": after all, I have confused other 'Z' words before!

What this amounts to is taking the conditional analysis I have given of 'qualia' and reading it as a kind of metasemantic account of what the semantics of 'qualia', 'Zombie', and so on would be in various circumstances.

There are reasons in general why I think *A*-intensions deserve to be regarded as giving the meanings of terms. After all, we are often ignorant of how the actual world is (consider our state in respect of 'water' in the fifteenth century) and yet manage to communicate something, and be—in some sense of competent—competent with the term.[35] I take it that if this conditional is right about what we would say and how we would behave on different suppositions about what is actual, it goes some of the way to explicating the competence we have when we are ignorant of exactly how things are, and what we communicate when we do communicate in such cases. While ignorant of the nature of water, when I ask for the water to be passed, it comes my way, even if my interlocutor is equally ignorant. Supposing that 'water is the actual watery stuff' exhausts their grasp of the term, explains the otherwise mysterious arrival of the water. Supposing that

[35] Jackson, "Why We Need *A*-intensions" (forthcoming).

the semantics are exhausted by water's connection to H_2O leaves that utterly mysterious. Now, it might be that there are historical contingencies. Perhaps if the Saxons had never invaded Britain, 'water' would have meant "foul foreign drink." But this is genuinely a mere metasemantic claim, for my use of the term does not depend on my opinions about British history. And similarly, the kinds of historical influences that could have made 'Zombie' mean "Zebra" are ones that do not regulate my use.

V.2. A further objection: the wrong conceivability has been explained. Chalmers has pressed a further objection. I have explained why it is conceivable that it is possible that there could be a world that has physical states of the required functional kind and no spooky states, and yet lacks qualia. But Chalmers thinks explaining this conceivability does only half the job. He thinks that what is equally important is that it is directly conceivable (and thus epistemically possible) that there are physical states of the right functional kind, no spooky states, and no qualia. In other words, for all that we know a priori, things might be such that there are functional states and no qualia.

This is, however, exactly what I deny. The conclusion in the section above is that the epistemic gap between physical states and qualia is a shadow of the conceivability of the possibility. It is not directly conceivable that there be functional states in the world, no spooky states, and no qualia, in part because on my view it is inconceivable that physical states could actually fail to be qualia if there are no spooky states. It seems to me that this is not a technical by-product of my account, but accounts for widely shared intuitions.

V.3. The argument from mere causal shift of meaning. There is another complaint of the same ilk. Presumably it is *possible* that there is some strange empirical fluke which ensures that the discovery that physicalism is false, say, would make me shift meaning so as to adopt a new concept 'qualia' and thus 'Zombie', such that it was right to say 'Zombies are possible'. Perhaps this has happened in history. Perhaps 'water' really does mean H_2O now, where it did not before. The discovery that water is H_2O has changed our concept due to the high salience of chemical nature. The objector goes on to say that we cannot tell the difference between the concept being the conditional one, and a metasemantic conditional changing the concept. Do we have a conditional concept, or a conditional about concepts?

But there is a difference, I think. There may well be all kinds of empirical discoveries that could change our conceptual palette. Sometimes we do change concept, and retain orthography, and maybe the concept 'water' is not the same as the concept that was around before the periodic table. But there is a test, at least in principle. Has history changed our concept, or are we still under the governance of the *A*-intension given by 'water is the actual watery stuff'? If the former, then 'water' just means H_2O, the concept has changed, and if we discover that actually there is no H_2O, we will have discovered that there is no water, and that is the claim we will make. I will mention only the beginnings of an account of governance here, but the idea will be that governance holds if after having changed, we are still disposed to change back if we revise back our empirical opinion. If the change would not be revised back should our empirical opinions change back, then governance does not hold. If we are still under the governance of the old *A*-intension, then discovering that there is no H_2O, but that (surprisingly!) the watery stuff is XYZ will make us claim that we were wrong about water being H_2O—in fact, it is XYZ. Further changes of view about the actual world—back to belief that the watery stuff is H_2O or on to other candidates—will always result in the claim that water is whatever we take to be the (actual) watery stuff. The difference between a metasemantic claim about what would change a meaning or replace a concept, and a conditional semantic analysis is just this kind of continued governance. And the fact that I take it that we (most of us) hold fixed that we have phenomenal experiences—if I become persuaded of dualism I think they are dualistic, if I lose the faith I decide they are physical, if I regain the faith I think they are dualistic, and so on—is what supports the idea that the conceptual structure is conditional. It is not that were I to become persuaded of dualism that I would change my concept, for I remain sensitive to the claim about the actual world, and would revise back if I became a physicalist. Of course, one *might* have a different concept from most of us, which was merely metasemantically conditional. Someone who was a physicalist about qualia, but on the discovery that dualism was true became a dualist about qualia, and then, on returning to his view that there were no spooky states, became not a physicalist about qualia again, but rather an eliminativist, would perhaps have such a concept.

VI. WHAT KIND OF ANALYSIS?

The benefit of the story so far is that it accounts for the intuition about the possibility of Zombies, which a straightforward analytical functionalism as a theory of the meaning of qualia language could not.

The difficulty is that the story is still intended to be a kind of analysis: the conditional is in some sense a priori and analytic. But it might not sound like it is something that is a plausible story in pure analysis. There are two methods that we have come to expect might deliver meanings. The first is empirical inquiry—the way discoveries about the actual nature of water guide us to its (C)-intension. The second is introspection—the way we figure out that bachelors are unmarried by consulting our own understanding, or the understanding of others if we suspect that our own is deficient.

The conditional analysis I have proposed is not supposed to be a posteriori. We do not look at the nature of actual qualia to find that the conditional analysis is true. It is, after all, an analysis.

But if it is right, neither is it based (solely) on self-evident introspection. So how can it be analytic? What might be the basis for the claim that something like this analysis is true?

The answer is that it describes how agents seem actually to behave. Many who in fact believe that phenomenal experiences give acquaintance with the essential nature of nonphysical properties or substances believe that phenomenal experience is essentially this. But in fact few people who once believed in such properties or substances, but became converts to physicalism, are in fact error theorists about phenomenal experience—rather, they flocked to naturalistic substitutes. Perhaps some of those who now believe in the dualistic properties would come to be error theorists, but it seems unlikely because the notion of phenomenal experience plays an indispensable role in our everyday communication about how the world is.

So what do we say to those who claim that not only do they believe in nonphysical properties, but they regard them as necessary for qualia? When they do the thought experiments, they insist, it is clear that spookless worlds are Zombie worlds. The answer is that the sort of evidence that underwrites these higher-order analyses is not only thought experiments; or at least not thought experiments that can be performed without particular experience. Thought experiments only tell us what we *think* we would say under certain circumstances. Some people might in fact believe in the nonphysical properties, and believe them to be essential to phenomenal experience. They might also think that they would become error theorists about phenomenal experience should they come to think, on metaphysical grounds, that there are no nonphysical properties. This latter thought could be one that they have access to via thought experiment: a kind of self-simulation. They imagine that they have become persuaded of the metaphysical error of dualism, and then wonder if this means they would persist in ascribing the intension they currently ascribe to "phenomenal properties," and thus they become error theorists. They do the thought experiment (roughly, the indicative version of the counterfactual thought experiment) and conclude that this would be the discovery that there would be nothing worth calling 'qualia'.

But they might be mistaken about this. This is why, at the most abstract level, thought experiments are an only fallible guide to analysis. The evidence, insofar as there is any, is that people are prone to change their views on the C-intension, rather than conclude that there was never anything that they were talking about when they used the word 'qualia'.

Why should this, if it is true, be victory to me? Why should I prioritize, in establishing the intension of a term for an agent, what he would in fact say in certain circumstances over the analysis he *explicitly* holds which is based on his best thought experiments concerning what he would say?

Something like a principle of deference to better-informed future time slices is what is at work here. If some future temporal stage of yourself were actually to find out that a certain metaphysics is true, and if there were no conspicuous loss of rationality attendant on this discovery, then your future self is better informed, that is, possessed of details of how things are that are elided in mere thought experiments. Why would it be reasonable to choose to go with opinions based on your pale attempts to self-simulate, rather than admit that while your self-simulations are the best guess you can give, what is right of you is determined by what you really *would* do when better informed. It is considerations like this that, I think, move those who think that when confronted by intelligent-seeming androids and so on, those in the thrall of John Searle would largely change their minds about the androids' unintelligence—and thus the term 'intelligent' in their mouths does not actually *mean* anything different from the term in the mouths of the AI community.

One last possibility remains. Perhaps an agent might not just have false beliefs about how he would behave in these circumstances. He might not *care* how he behaves or uses language in conditions of superior information. He might explicitly announce that the terms he uses definitely have the intentions he takes them to have, making it explicit that he is not bound by how his better-informed, and by some lights, reasonable future time slices might think or behave. Call this view *analytical Muggletonianism*. The Muggeltonians were a seventeenth-century English religious sect who believed that reason led to evil and error.[36] Analytical Muggletonians do not defer to their future better-informed or more reasonable time slices.[57] For them, current thought experiments and so forth are the whole of what determines the intentions of their terms, and they regard their future time slices, insofar as they disagree with current opinion, as mistaken—or as having radically changed the subject without realizing it.

---

[36] Richard Holton introduced the Muggeltonians to philosophical discourse in his "Reason, Value, and the Muggletonians," *Australasian Journal of Philosophy*, LXXIV (1996): 484–87.

Could there be analytical Muggletonians? I am happy to accept that there could be. Others may make some constitutive claim that the Muggletonian, for all his self-binding stipulation, is just mistaken. I see no reason to deny that we can and do define things stipulatively. But to the extent that analytical Muggletonianism is mere stipulation it seems hardly a move that will help resist my analysis. For anyone who is not a Muggletonian and discusses qualia wants to be using the term that the rest of us use when 'qualia' proceeds from our mouths or keyboards. And to the extent that most of us find normative reason to defer to our better-informed future selves (or our better-informed colleagues in the present), we are using the term differently from the Muggletonian.

All of this may appear to paint a slightly different picture of the analytic than is traditional. The analytic truths supervene on psychology: that much is preserved. But they do not supervene on the immediately introspectible part of psychology, or even on rational progress from the introspectible parts of psychology. For no amount of armchair speculation can deliver the judgments that would in fact be delivered in real circumstances. My acceptance of the possibility of analytic Muggletonianism means, however, that there is still an introspectible component that governs meaning: it is the deference that most of us do have to future better-informed opinion. We defer to our own better-informed counterparts in much the same way as, on other matters, we defer to other, better-informed, individuals. It is in virtue of this that what we mean by 'qualia' is still, however indirectly, in our own control.[58]

---

[57] You do not get to be an analytical Muggeltonian just because there are some better-informed future time slices to whom you do not defer, for it could be perfectly reasonable to be wary of the semantic intuitions of some future time slices of yourself overwhelmed by too much information: hence the 'by some lights reasonable' clause.

[58] So it should be clear that nothing in what I say about the two-dimensional framework assumes that the framework itself removes the need for an account of what an analysis is. I agree with Block and Stalnaker (*op. cit.*) that two-dimensionalism only formalizes the conceptual component of meaning, it does not provide it. The claim that the conditional analysis of qualia I give here is correct is supposed to be independently justified by arguments like the ones in this section. The role of the two-dimensional apparatus in this paper has been to show how such an analysis undercuts the arguments from conceivability.

VII. SPECULATIVE METAPHYSICS

How do we determine the truth or falsity of physicalism? For this is something that we need to know if we are to determine the C-intension claims involving terms like 'qualia', or to establish the (superficial in Evans's sense)[30] possibility or otherwise of Zombies. Analysis is of no help with this task. Just as the analytical functionalist has to appeal to other considerations to determine whether the states that play the causal roles are physical or not, the kind of conditional functionalism I advocate here requires the metaphysics of the mind to be settled by some other method than analysis. Analysis comes after speculative metaphysics. There is an interesting question, perhaps outside the scope of this article, as to what kind of endeavor that is. For it does not seem to be the kind of thing we can settle a priori, and yet it seems to be beyond empirical confirmation as well. Still, this is not a problem that is local to philosophy of mind. Any robust realism allows that there are questions like this. Are there spatiotemporally and causally isolated parts of actuality? Experience will not settle it—since if there are, they are causally isolated. And yet surely it is not a priori either. Perhaps it is a posteriori for some ideal agents. The challenge for philosophers is why we think there are even plausibility arguments either way on such questions.

One final worry in this direction: What if I *do* think that the truth or otherwise of dualism is an a priori one? Suppose that my suspicion is that the idea of a property whose intrinsic nature is available to experience is somehow incoherent (and on alternate Tuesdays this seems right to me). Can I still hold an *analysis* of qualia which is conditional and that has claims about the truth or otherwise of physicalism in its antecedent?

I think so, but it will get messy. There is some good sense of epistemic possibility on which when I worry about the coherence of the spooky state, it is epistemically possible either way. But of course we cannot then regard the worlds considered as actual as being possible worlds *simpliciter*. For on this assumption, there may be no possible world with the spooky states. Thus, we will have to regard the rows in a two-dimensional matrix other than the row containing the actual world, as containing different sorts of items—not logically possible worlds, but epistemic possibilities of a certain kind.

VII.1. Qualia and consciousness. One final clarification: so far I have talked about qualia throughout. But the argument for dualism from the possibility of Zombies is often thought of as depending on issues about consciousness. Elsewhere,[40] I have argued that we get consciousness for free when we mix qualia with other mental states where the issue of physicalism is more easily settled. Anyone who accepts that can go away happy. Anyone who does not has an 'elementary exercise for the reader' to perform: she must reconstruct the story so far in terms of consciousness.

---

[30] See Davies and Humberstone.

VIII. CONCLUSION

So a powerful argument for dualism is defused. The Zombie intuition can be taken as a guide to meaning, and yet not with the result that we must abandon physicalism. This is because the intuition is the product of a complex conditional analysis which I claim is true of 'quale', and when we see what part of the meaning of the term gives rise to the intuition, we see that it is metaphysically idle. Of course, this does not settle the issue of physicalism the other way. Speculative metaphysics lives. But many of us as are convinced as we ever were of the plausibility of physicalism. And on the conditional analysis that I offer, if we are right about this metaphysical claim, then a substantive functionalism is true about qualia: they just are states that realize certain functional roles.

DAVID BRADDON-MITCHELL

University of Sydney

---

[40] Braddon-Mitchell and Jackson, *op. cit.*, p. 135ff.

THE SCOPE ARGUMENT[*]

In his classic essay, "Logic and Conversation," H. Paul Grice[1] sets out to demonstrate that certain formal devices of logic, such as sentential connectives and quantifiers, have the same meaning as their natural language counterparts. Contrary appearances had spawned efforts in natural language semantics to marginalize formal logic, on the one side, and "correct" natural language, on the other. Since these efforts assume that there are divergences in meaning, however, Grice regards them as wrongheaded. Using implicatures, he argues that we can explain these appearances while remaining committed to the interpretations these devices are given in formal logic. Grice's efforts did not go unrewarded: his implicature-based approach to the semantics and pragmatics of conversation became the standard. Even so, people on both sides of this dispute have challenged Grice's approach on many fronts. Prominent among these is a challenge first issued in L. Cohen[2] and later taken up by Robyn Carston[3] and François Récanati,[4] among others. These critics argue that certain data cut against Grice's approach, namely, complex examples in which sentences with a specific type of semantics are embedded inside the scope of sentential connectives. Grice endorses a truth-functional interpretation of the sentential connectives, a move that seems justifiable for a wide range of cases; however, with respect to these complex examples, this commitment ostensibly forces him to see semantic problems where intuitively there are none. Moreover, certain examples in which connectives are embedded inside the scope of other connectives appear to undermine entirely the Gricean interpretation of these terms. It is urged that any attempt to respond

---

[*] I am grateful to many for commentary and criticism. The following people deserve special mention: Kent Bach, Anne Bezuidenhout, Eros Corazza, Bruce Glymour, Ernie Lepore, John Perry, François Récanati, Stephen Schiffer, Corey Washington, and audience members at a UI/WSU Philosophy Colloquium. I am especially indebted to Joseph Keim Campbell, David Shier, and Harry Silverstein, as they encouraged me to write this article and gave me helpful advice along the way.

[1] In *Studies in the Way of Words* (Cambridge: Harvard, 1989), pp. 22–40.

[2] "Some Remarks on Grice's Views about the Logical Particles of Natural Language," in Yehoshua Bar-Hillel, ed., *Pragmatics of Natural Languages* (New York: Humanities, 1971), pp. 50–68.

[3] "Implicature, Explicature, and Truth-Theoretic Semantics," in Steven Davis, ed., *Pragmatics: A Reader* (New York: Oxford, 1991), pp. 33–51.

[4] "The Pragmatics of What Is Said," in Davis, ed , pp. 97–120; see especially pp. 112–14.

0022-362X/03/0003/136–57

to these examples with implicatures fails, and so Grice's account is incapable of explaining the range of cases for which it is intended. Therefore, it must be set aside in favor of an alternative account of the semantics and pragmatics of conversation.

I believe that this argument, which I will call *the scope argument* (TSA),[5] is unsound, and in this essay I craft a Gricean rejoinder. In developing TSA, I focus on a particular example first introduced in Cohen, an example that supports what I take to be the most threatening version of the argument. After close analysis of this example, I present a detailed statement of the version of TSA that it supports. This argument is intended to demonstrate that Grice cannot retain his semantic commitments and account for the interplay of meanings in communication, but it fails on two fronts. First, it underestimates the resources available to the Gricean for dealing with complex examples of this sort. Second, and more importantly, it misrepresents the nature of Grice's implicature-based account, evaluating it as if it were a contribution to the psychology of conversation. This treatment reflects a failure to appreciate the specific character of Grice's contribution to the theory of conversation, a failure that vitiates TSA.

I. THE SCOPE ARGUMENT

In his semantics, Grice interprets sentential connectives truth-functionally, shifting non-truth-functional aspects of meaning to pragmatics where they are modeled with conversational implicatures. In developing this theory, he uses many examples, but critics have argued that these are too limited, and that when one considers a broader range, the intuitive appeal of his theory dissipates. Cohen, for instance, remarks, "[Gricean] theories gain what support they seem to have from the consideration of relatively simple examples. Their weakness becomes apparent when more complex sentences are examined..." (*op. cit.*, p. 67). As an instance of such a datum, he supplies a sentence quite like this one:

(i) If the old king dies and a republic is declared, Tom will be content, and if a republic is declared and the old king dies, Tom will not be content.[6]

This sentence is formed by embedding complex sentences that contain sentential connectives inside the scope of other sentential connectives. In fact, there are two layers of embedding: the principal connective, namely, the second 'and', joins two complex conditionals that contain conjunctions as their antecedents. If the connectives are treated truth-functionally and no unarticulated semantic structures are posited, an utterance of (i) cannot be true if the sentences conjoined in the antecedents of the embedded conditionals are both true. But surely, Cohen argues, we can and do regard this sentence as expressing a proposition that could be true in just such a case. Thus, there would appear to be a conflict between the Gricean take and our intuitive take on the semantic content of these utterances.[7]

---

[5] This name is inspired by Récanati, who addresses these considerations while discussing what he calls the "scope principle" It is important to keep TSA and Récanati's scope principle distinct, however. The scope principle is introduced as a universal test for semantic status, while TSA is designed to undermine the implicature approach of Grice through provision of a counterexample.

[6] Adapted from Carston, p. 45. Carston follows an example on p. 58 of Cohen.

Intuition enjoins us to see (i) as nonproblematic, but the Gricean commitment to truth-functional interpretation of the semantics of connectives suggests otherwise. This can be cast as a conflict between two premises. First, our *intuition premise*:

(P1) An utterance of (i) is intuitively nonproblematic.

And, second, our *minimalist premise*:

(P2) The semantics of natural-language connectives is given entirely by their standard truth-functional interpretation.

Each of these premises deserves comment. We can begin by remarking on what it is for an uttered sentence to be *intuitively nonproblematic*. Emphasis on intuition here makes plain the importance of how an utterance *strikes* us. As fluent speakers of English, our immediate reaction to an utterance of (i) determines whether it is intuitively nonproblematic or not.[8] But in general, what is it for an utterance to be intuitively nonproblematic? For our purposes, an utterance will be intuitively nonproblematic just in case it is taken to be a vehicle of a substantive and possibly true claim. Consider (i). It can be used to make a true claim if the conjunctions that form the antecedents of the embedded conditionals are false, but this would be a trivial and not a substantive claim. Only in those circumstances where these are true could (i) be a vehicle of a substantive claim. To say that an utterance of (i) is a "vehicle" of such a claim might only be to say that it is used to convey or imply it comfortably, given the conventional meanings of its terms in English. Thus, even if it cannot be assigned a substantive and true claim as its conventional interpretation, it could count as intuitively nonproblematic in a circumstance if it could be taken to imply one there. What really matters here is that one rather effortlessly interprets the utterance as conveying or implying a substantive and possibly true claim, regardless of the specific formal relationship between the utterance and the claim. Finally, taking an utterance as intuitively nonproblematic reflects the perspective of one who "takes" it; this is meant to capture the fact that while one could intend an utterance to be intuitively nonproblematic, whether it is or not will depend on how it strikes those who attend to it.

---

[7] This conflict can be generated by other data as well For examples, see Cohen, p. 63; Carston, p. 45; and p. 203 in Anne Bezuidenhout, "The Communication of *De Re* Thoughts," *Noûs*, xxxi (1997): 197–225. In general, the data appear to vary along two dimensions. First, there is the dimension corresponding to the sentence embedded, which either contains a connective, such as 'and' or 'if .then', or a predicate, such as 'jumps', associated with enrichments that do not fill syntactic slots. Second, there is the context within which the sentences are embedded. Two types of context are represented in the data: (a) anti-reflexive operators, such as *better than*, and (b) binary sentential connectives that have the following property: when the first sentence connected is true, the truth value of the complex sentence formed depends on the truth value of the second sentence. With respect to the second context, the conjunction and conditional exhibit the associated property, and they appear regularly in examples found in the literature. By contrast, the disjunction does not have this property and so does not create a context for the purposes of TSA. It would appear that any combination of embedded sentence type and context type produces scope argument data, although this essay is not the place for that argument.

As for the minimalist premise, let us begin by considering the motivation behind it. Among other things, natural language is a medium for stable and robust inferences. One can explain this fact about language by taking a cue from formal logic and interpreting the inferential elements (for example, the sentential connectives) truth-functionally. This would supply the systematicity and generality necessary to underwrite the inferential character of natural language as it is used in communication. In addition, it would help explain the stable contributions made by terms and sentences to communication across a wide variety of communicative contexts. Explanation and prediction across these contexts require a general foundation and this proposal fits the bill, specifying as it does those properties that account for the truth-bearing and truth-conducting structure of linguistic elements.[9]

---

[8] While not as big a fan as some, Grice agrees that intuitions play an important role here—see Grice, p. 49. For a bigger fan, see Récanati, *Direct Reference* (Cambridge: Blackwell, 1993). Cf. Kent Bach, "Seemingly Semantic Intuitions," in Joseph Keim Campbell, Michael O'Rourke, and David Shier, eds., *Meaning and Truth* (New York: Seven Bridges, 2002), pp. 21–33.

[9] See the discussion of the "conversationalist hypothesis" in Cohen, pp. 50–52. Cohen is careful to avoid attributing this view to Grice, although there are more than a few indications in Part I of Grice that he would welcome such an attribution. Cf. Jens S. Allwood, "Logic and Spoken Interaction," in Terry Myers, Keith Brown,

The general attitude motivating (P2) also motivates a minimalist approach to the semantics of natural language, and this is certainly evident in Grice. He uses 'what is said' to refer to the semantic core that comprises the conventionally encoded meanings of sentential constituents, arranged in an order determined by the syntax of the sentence, as well as those contextual determinants necessary to disambiguate and fix indexical elements, that is, those determinants necessary to make the utterance truth evaluable. Any element of meaning associated with a sentence that is not on this short list counts as *pragmatic content* and so is not a part of what is said.[10] Commitment to (P2) implies that sentential connectives contribute only their bivalent truth-functional meanings to what is said by an utterance. (In what follows, I will treat the term 'semantic content' as synonymous with 'what is said'.)

With (P1) and (P2) so understood, the conflict between the premises generated by (i) is even more evident. If we limit ourselves to the truth-functional meanings of the connectives in (i), we cannot make it express a substantive and truthful claim, in violation of (P1). It can express such a claim, however; thus, it would appear that Grice's commitment to (P2) is mistaken. At this point, though, Grice would turn to his implicature machinery, which is designed to do justice to (P1) while retaining (P2). From his perspective, the conflict between these is merely apparent and rests on an altogether too simple view of the matter. We need not reject either premise so long as we construe (P1) properly. The apparent tension is resolved if we allow that our intuitions about the significance of an utterance might not be all that discriminating. Intuitions can indicate when a true proposition figures prominently into the overall interpretation of an utterance; however, they cannot in general determine whether it forms a part of the utterance's semantic content or a part of its pragmatic content. Thus, when our intuitions tell us that examples like (i) are nonproblematic, they pass judgment on the total content of the utterances, that is, their semantic content together with their pragmatic content, or what Grice calls their "total significance" (*op. cit.*, p. 41). Given that our

---

and Brendan McGonigle, eds., *Reasoning and Discourse Processes* (San Diego: Academic, 1986), pp. 67–94.

[10] For related distinctions, see *semantic meaning/speaker meaning* in Saul Kripke, "Speaker's Reference and Semantic Reference," in Peter A. French, Theodore E. Uehling, and Howard K. Wettstein, eds., *Contemporary Perspectives in the Philosophy of Language* (Minneapolis: Minnesota UP, 1979), pp. 6–27. See also *what is said/what is implicated* in Recanati, *Direct Reference*, and *implicature/explicature* in Dan Sperber and Deirdre Wilson, *Relevance: Communication and Cognition*, Second Edition (Cambridge: Blackwell, 1995), p. 182.

intuitions do not discriminate between semantic and pragmatic content, it is open to Grice to locate theoretical reasons for discriminating them; in particular, he can assert (P2), so long as the rest of the content necessary to explain the intuitively nonproblematic character of these utterances is accounted for in pragmatic terms. This is precisely what he does: he embraces (P2) and then uses the machinery of conversational implicature to account for satisfaction of (P1) in these cases. They seem nonproblematic because we effortlessly resolve the tension between their minimalist semantics and intuition by identifying the conversational implicatures that dominate our interpretations. Thus, Grice endorses the following premise:

> (P3) One can maintain (P1) and (P2) if one introduces pragmatic elements, and specifically conversational implicatures, into the interpretation of utterances.

This move is intended to be general, and it would appear to account for a variety of examples; however, Cohen argues that it does not work for an example like (i), which therefore counts as a counterexample to the Gricean approach. Armed with (i), he mounts a reductio of the Gricean view, arguing that any attempt to restore (P1) with pragmatic machinery, and conversational implicatures in particular, undermines a commitment to (P2).[11] To account for our intuitive interpretation of (i), Cohen notes that the Gricean must introduce conversational implicatures that express the temporal sequence associated with the embedded conjunctions. Grice's conversational maxims support these implicatures as additional meanings that can explain away the apparent conflict between an utterance of (i) and the cooperative principle. In particular, the maxims of quality and manner support calculation of a temporal solution to the apparent truth-conditional problem generated by the semantic content of (i), namely, that if interpreted truth-functionally, the conjunctive antecedents would make it impossible to take an utterance of this sentence to convey a substantive and true claim, a violation of the cooperative principle in a normal conversational setting. Thus, it would seem that identification of semantic content occurs prior to identification of implicatures, since the former causes the problem solved by the latter. In identifying implicatures, we treat what is said by a sentence as a fully formed object that can be evaluated in light of our conversational expectations. Call this way of working out pragmatic content the *serial generation approach* (SGA), since we identify the semantic content and then the pragmatic content in series. Proponents of TSA regard SGA as part of the Gricean solution.[12] Thus, we have our fourth premise:

---

[11] My reconstruction of this part of the argument is grounded in the reasoning found in Cohen, pp 58–59 and pp. 63–64. See also pp. 194–95 in R. Posner, "Semantics and Pragmatics of Sentence Connectives in Natural Language," in John R. Searle, Ferenc Kiefer, and Manfred Bierwisch, eds., *Speech Act Theory and Pragmatics* (Boston: Reidel, 1980), pp. 169–203.

(P4) If (P3), then we generate pragmatic content in conformity with SGA.

Given SGA, we take what is said by a sentence to cause the conversational problems solved by implicatures. We solve these problems by introducing the pragmatic content into the total content of the sentence uttered. Thus, implicatures are associated with a sentence, but there is no requirement that this be the top-level embedding sentence. In an example like (i), there are a number of embedded sentences with which we might associate the implicatures. It is reasonable to expect some correlation between the sentences that underwrite implicature generation and those with which we associate the implicatures generated; however, this expectation is insufficient by itself to determine whether it is the whole sentence or just some sentential part of the whole that will be associated with the implicature. Thus, we appear to have two options for association: (a) we can associate the implicature with the top-level embedding sentence, which in this case is the conjunctive sentence formed out of the two conditionals, or (b) we can associate it with an embedded sentence, either the simple sentences, the conjunctions embedded in the antecedents of the conditionals, or the embedded conditionals themselves. Call the first of these options *holistic* and the second *atomistic*. Thus, we have two more premises of TSA:

(P5) If SGA, then we associate the pragmatic content with a sentence.
(P6) If we associate the pragmatic content with a complex sentence like (i), then we must associate it either holistically or atomistically.

---

[12] This assumption is central to the serial model of linguistic interpretation that is criticized throughout Récanati, "The Alleged Priority of Literal Interpretation," *Cognitive Science*, XIX (1995): 207–32. In this essay, he presents the case for regarding this as the *standard* interpretation of Gricean pragmatics. In addition, see Récanati, *Direct Reference*, pp. 271–72. In setting up his scope principle, Récanati endorses an argument that he attributes to Jean-Claude Anscombre and Oswald Ducrot, and in this he includes as a step the claim that "no implicature can be generated at the sub-locutionary level, that is at the level of an unasserted clause such as the antecedent of a conditional," p. 272. Rather, implicatures can only be generated at the level of the "complete utterance," which must therefore be given if the implicatures are to be worked out. See Anscombre and Ducrot, "Échelles Argumentatives, Échelles Implicatives et Lois de Discours," *Semantikos*, II (1978): 43–67.

Beginning with holistic association, note that SGA requires us to interpret the entire embedding sentence semantically before pragmatic supplementation. After calculating what is said by the utterance, we would be compelled by a clash with Grice's cooperative principle to search for implicatures. This implies, however, that we would detect a problem with the utterance that would require resolution by implicatures. If there were such a problem, then it would be reflected in our intuitive reaction to these examples, and it is not, *ex hypothesi.* Therefore, we must not associate the pragmatic content needed to resolve the problems generated by (i) holistically. Thus, we can add these steps to the argument:

(P7) If holistic, we would generate pragmatic content to solve an intuitive problem with (i).
(P8) (i) is intuitively nonproblematic, per (P1).
(9) Therefore, we do not associate pragmatic content with (i) holistically.

Thus, we are left with atomistic association. On this approach, we would associate implicatures with constituent sentences. Given the reasoning above, this cannot be done to solve intuitive problems, since there are none. Still, though, the utterance must give us a reason to generate implicatures. It is here that the atomistic approach might supply precisely what the Gricean needs. Perhaps implicature generation involves constituent sentences and occurs so quickly and effortlessly at the level of constituent sentences that the process is practically unnoticeable. If this is the case, then the atomistic approach would get you around obvious clashes with (P1) at the level of the whole sentence by distributing clashes out over the constituent parts, each of which would be resolved quickly and so be too small and ephemeral to receive much notice.

Whether this is sustainable or not, a knotty problem remains. To bring this out, it will help to regiment the logical form of (i) as follows:

(ii) $((S_1 \& S_2) \rightarrow S_3) \& ((S_2 \& S_1) \rightarrow -S_3)$

There are three points at which we could attach implicatures, consistent with the atomistic approach: (a) the simplest sentences (that is, S_1, S_2, and S_3), (b) the conjunctive sentences that form the antecedents of the conditionals, or (c) the conditional sentences. Option (a) seems unlikely, since the complex sentences are what create the problem. That leaves options (b) and (c). If we associate the conjunctions '$S_1 \& S_2$' and '$S_2 \& S_1$' with implicatures that express their temporal ordering, then their implicative contents will serve as inputs into computation of the truth values of the conditionals. This implies, however, that the truth values of the conditionals will depend on more than just the truth values of their constituent sentences, a fact that conflicts with (P2). Similarly, if we attempt to resolve the difficulty by associating implicatures with the conditionals, we would be forced to treat the principal conjunction non-truth-functionally. Thus, the atomistic attempt to capture our intuitions, and so satisfy (P1), forces us to renounce (P2). To our argument, then, we add these steps:

(P10) If atomistic, we must treat connectives in (i) non-truth-functionally.
(P11) They must be treated truth-functionally, per (P2).
(12) Therefore, we do not associate pragmatic content with (i) atomistically.

Since (P6) offers no third alternative, we infer that SGA must be false; this, however, implies that we must reject (P3), and so we have our next step:

(13) Thus, the pragmatic analysis involving implicatures in (P3) will not resolve the problems raised by an example like (i).

This leaves us without a viable way of using conversational implicatures to maintain both (P1) and (P2) in the face of (i). Thus deprived of implicatures and bereft of an alternative, we are forced to relinquish our grip on either (P1) or (P2).[13] Since (P1) is the more compelling of the two, (P2) is shown the door, and with it goes the Gricean view of the semantics of connectives.

---

[13] Opinions about what to do here vary. First, we might reject (P1) and claim that the examples are really problematic after all, despite intuitions to the contrary. For us to reject this prima facie case in favor of (P1), however, there would need to be significant theoretical benefit associated with commitment to (P2), and that has yet to be established

Second, we could reject (P2). This option comes in several different varieties that we can divide into two groups: the semantic and the pragmatic. The semantic varieties include positing ambiguity (see Grice, pp. 47–49 for critical discussion), introducing slots to be filled (see Récanati, "The Pragmatics of What Is Said," p. 103 for discussion), and introducing features to be killed (Posner, pp. 182–88). The pragmatic varieties view much of the content that goes beyond (P2) as "free enrichments," determined pragmatically in unpredictable ways by the interaction of utterance and context, and added to the semantic content of utterances as directed by our intuitions (see, for example, Bezuidenhout; Carston; and Récanati, "The Pragmatics of What Is Said"; see also Sperber and Wilson, especially p. 183). I believe, however, that it is too early to give up on a Gricean approach in favor of one of these alternatives. In what follows, I will argue that upon closer inspection of the data in question, we can retain both (P1) and (P2).

(14) Therefore, we must reject (P2) and the Gricean view of the semantics of connectives along with it.[14]

II. A GRICEAN REJOINDER: PART ONE

The Gricean believes that one can and should maintain the following theses at once:

(A) Utterance of (1) is intuitively nonproblematic.
(B) The semantics of sentential connectives is exhaustively given by their truth tables.
(C) We can solve problems of interpretation at the semantic level by introducing content at the pragmatic level.

Call these his *central tenets*. Thesis (A) is a starting point for all who participate in this debate. Thesis (B) is what sets the Gricean view apart. Commitment to truth-functional interpretation of connectives as the stuff of semantics reveals an even more basic commitment to the existence of an objective level of utterance significance that supports robust communicative exchange. As such, (B) springs from the belief that semantic minimalism supplies the best framework within which to examine fundamental and very general aspects of utterance significance, namely, those associated with sentences that support variable contextual interpretations without being themselves sensitive to these variations. The Gricean requires (C) to smooth over prima facie tensions between (A) and (B). Thesis (C) is needed to capture those aspects of utterance significance that *are* sensitive to contextual variation.

---

[14] Variations of this reasoning are found in the literature. In all cases, the data are regarded as counterexamples to Grice and support variants of TSA. In some cases (for example, Posner), the supporting examples are similar to those of Cohen, whereas other cases build on different types of recalcitrant examples. For instance, consider this sentence:

(iii) It is better to meet the love of your life and get married than to get married and meet the love of your life (from Wilson, as reported in Carston, p. 45).

Examples like this motivate a form of reasoning that differs from the argument regimented in the text by ignoring the atomistic approach and relying exclusively on appeal to intuition. For discussion, see Bezuidenhout; Carston; and Récanati, "The Pragmatics of What Is Said." In this version, there is no use made of (P4). Instead, the argument proceeds from (P3) in the following fashion:

(P4′) Use of implicatures to account for the significance of examples such as (iii) requires recognition of a problem with the example.
(P5′) There is no problem with the example, per (P1).
(6′) Therefore, we must reject (P3).

This form of the argument is arguably weaker than the Cohen formulation, as it depends solely on an appeal to intuition, and intuitions about these matters can differ in ways that are difficult to adjudicate. The form of the argument reconstructed in the text demonstrates that intuitive dissonance, assumed for the sake of argument, prompts a response that results in truth-conditional distress. Thus, it works as an internal criticism of Grice, whereas the form described in this note is primarily external.

Whatever proponents of TSA may think of the Gricean commitment to objective meaning, they are of one mind in maintaining that the Gricean view is untenable. As we saw in the previous section, they believe examples like (i) demonstrate that commitment to (B) threatens (A), the most intuitively plausible thesis, and that any attempt to preserve (A) with (C) undermines (B). Thus, (C) will not help us save (A). Failing any other Gricean alternative, we are left with the option of rejecting (A) or (B), and (B) loses. Therefore, they argue that the Gricean view of the semantics of connectives is mistaken and should be set aside, a move that would undermine the Gricean approach to conversation as a whole.

As we will see, TSA takes very seriously the processing implications of pragmatics. Pragmatics concerns aspects of linguistic significance that are borne out of the intentional use of language, and Grice's view is a seminal contribution to this discipline. Proponents of TSA note the emphasis Grice places on calculation and generation in his story, which they believe signals his intention that it be a partial explanation of the processing of pragmatic content in utterance interpretation. In light of this, they believe they are justified in bringing processing considerations to bear on Grice's view.[15] As they see it, if they can establish that there is an irreconcilable conflict between utterance processing and his pragmatic theory, then his theory will have to go.

I will argue in the next section that the Gricean view should not be taken in this way, but whatever one thinks about the relationship between Grice's pragmatic story and utterance processing, it is clear that his account of the logic of conversation must at the very least be *compatible* with the psychology of conversation if it is to be adequate. Indeed, if the psychological evidence established a processing model that was inconsistent with his account, then so much the worse for his account. As I have characterized it, TSA urges precisely this conclusion by pressing (i) as a counterexample to Grice's view of conversation. In particular, it serves (i) up as proof that we cannot generally process utterances in the fashion described by Grice while remaining fast to his view about the meanings of sentential connectives. This only follows, however, if we embrace (P4), that is, if we accept that Grice's implicature approach implies SGA. It is perhaps worth noting that there is no compelling textual evidence for attributing this premise to Grice, even though this is the standard interpretation. But whatever Grice thought, it is not the case that his central tenets require him to endorse it. In this section, I lay out an alternative to SGA that involves both on-line and parallel processes. I argue that this alternative is compatible with Grice's central tenets, and that this compatibility undermines TSA.

---

[15] This is apparent in Bezuidenhout, Récanati, "The Pragmatics of What Is Said," and *Direct Reference*; Carston; Sperber and Wilson; and even in Cohen. It is also a widespread attitude in cognitive psychology. For a list of psychological works that rest their criticism on this assumption, see pp. 51–52 in E. Wisniewski, "The Psychology of Intuition," in Michael DePaul and William Ramsey, eds., *Rethinking Intuition* (Lanham, MD: Rowman and Littlefield, 1998), pp. 45–58.

The content communicated by an utterance of (i) can be distinguished into semantic and pragmatic aspects. Focusing on the connectives, a Gricean will take their semantics to consist in their truth-functional character, which is a noncancelable aspect. Their pragmatic aspects, by contrast, are not truth-functional, and they are cancelable and calculable from the conversational maxims.[16] These aspects are closely related to specific terms, such as 'and' and 'if'. Consider the first 'and'. In addition to the explicit conjunctive element, this logical particle contributes the temporal element 'then', which is cancelable and calculable and so qualifies as pragmatic content.[17] Similar stories can be told about the pragmatic content associated with the conditionals and the 'and' that occurs in the antecedent of the second conditional.[18]

TSA assumes a conceptual connection between use of implicatures to account for pragmatic aspects of content and SGA. If in fact the maxims are used to calculate the pragmatic content, they would appear to require the whole utterance on which to operate. There is nothing, however, about the Gricean picture of content just sketched that requires all pragmatic elements to be calculated relative to the maxims; to be sure, they must be *calculable*, but they need not actually be calculated.[19] Thus, how they come to be associated with an utterance in a given case can vary, and while it could be via an explicit maxim-driven calculation, it need not be. This is a good thing, since in many cases the pragmatic elements seem to come along with the sentence uttered. This close association could be due to a process of standardization, or perhaps through the standardized application of general pragmatic heuristics.[20] These elements would count as pragmatic in the relevant sense—that is, they are cancelable and calculable—even though they are not calculated. Further, there would appear to be no reason why the process of standardization cannot associate the content with subsentential chunks, such as words or phrases. As we noted, the pragmatic elements required to handle (i) are related to specific terms, and this relationship could be cashed out in terms of standardization. Thus, pragmatic aspects of content could come to an utterance context as prepackaged associations forged by their regular conjunction with a term. These pragmatic associations underwrite interpretive expectations and tendencies that shape the interpretation assigned to an utterance by conversational participants. Since we bring these tendencies with us to conversations, they enable us to give shape to interpretations while utterances unfold. Thus, we are conditioned through conversational experience to expect certain pragmatic associations, and so we look to save cognitive resources by assigning them immediately, all the while allowing that this assignment is defeasible and could be cancelled explicitly or contextually. Consistently with Grice's account, then, we can embrace a dynamic, *on-line processing approach* (OLPA), according to which the pragmatic content of a sentence is generated "on the fly" while the sentence is uttered.[21]

---

[16] Grice, pp. 39–40; and 43–44. Calculability is emphasized on p. 31.

[17] We could, for instance, explicitly reject the temporal implication as an aside while uttering (i), thereby establishing it as cancelable. As for its calculability, see section II

[18] For the purpose of countering TSA, we really need only cancelable and calculable content. Grice would consider these to be generalized conversational implicatures. For a defense of this classification, see Stephen C. Levinson, *Presumptive Meanings* (Cambridge: MIT, 2000), especially chapter 2. Others regard this classification as mistaken, as Bach argues in "Conversational Impliciture," *Mind and Language*, IX (1994): 124–62. As Bach sees it, these elements qualify as *implicitures*, or implicit content, as distinct from implicatures. For our purposes, it makes no difference whether these are generalized conversational implicatures or implicitures since both are cancelable and calculable and so fall into the domain carved out by the notion of pragmatic content we are using.

---

[19] See Grice, p. 21, where he emphasizes that implicatures must be "capable of being worked out."

[20] For discussion of standardization, see Bach, "Standardization and Conventionalization," *Linguistics and Philosophy*, XVIII (1995): 677–86. For discussion of this process as well as general heuristics, see Levinson.

[21] Words and phrases function as anchors for this pragmatic content, which operates as a rapidly accessed pragmatic presupposition that shapes ongoing interpretive efforts and creates expectations about the trajectory of the interpretation as a whole. See C. Martin, "Pragmatic Interpretation and Ambiguity," in *Program of the Eleventh Annual Conference of the Cognitive Science Society* (Hillsdale, NJ · Lawrence Erlbaum, 1989), pp. 74–481; C. Riesbeck and R. Schank, "Comprehension by Computer: Expectation-Based Analysis of Sentences in Context," in Willem J. M. Levelt and Giovanni B. Flores d'Arcais, eds., *Studies in the Perception of Language* (New York: John-Wiley, 1978), pp. 247–94, J Sedivy, M. Tanenhaus, C. Chambers, and G. Carlson, "Achieving

Consider (i) once again. As it is uttered, we hear an 'if' and we expect a conditional claim; further, we expect the antecedent and consequent to be connected in ways that are not *merely* truth-functional, unless our expectations are cancelled in some fashion by the surrounding conversational context. The 'if' in (i) will make the listener sensitive to a range of antecedent-consequent relationships, for example, causal, modal, mereological, and so forth. When the consequent appears, the listener will likely opt for a causal interpretation. Similarly, when we hear an 'and', we expect a second conjunct, and further, we expect the order of the conjuncts to be significant. With respect to the first 'and', the expectation is that it will be expanded into 'and then' because the conjuncts are events whose temporal relationship is often significant.[22] Therefore, introduction of prepackaged associations into the pragmatic mix makes possible recognition of the dynamic and on-line character of utterance interpretation, a fact that undermines (P4). As a consequence, TSA is rendered unsound.

Before celebrating this result, the Gricean must attend to a further problem that cannot be resolved by on-line processing alone. However we process the pragmatic content, holding fast to (P2) commits us to the view that the full semantic content of (i) cannot be true when both conjuncts in the antecedents of the conditionals are true. On the assumption that the semantic content influences our intuitive reaction to an utterance, this implies that one would encounter an intuitive problem with (i), a consequence that undermines (P1). Granted, this problem likely will not arise until on-line processing is complete, but at that point we would be left with semantic content that has the wrong truth-conditional profile, creating a problem for our interpretation. OLPA promises relief by eliminating SGA and its associated atomistic/holistic dichotomy, but it leaves us with a problem that is very much like the one created by the holistic resolution, namely, a conflict between intuitive interpretation and semantic con-

---

Incremental Semantic Interpretation through Contextual Representation," *Cognition*, LXXI (1999): 109–47; and I. Zukerman, "Expectation Verification: A Mechanism for the Generation of Meta Comments," in *Program of the Eleventh Annual Conference of the Cognitive Science Society*, pp 498–505. It should be noted that arguments such as those advanced by C. Clifton and F. Ferreira are relevant to determining the precise character of the pragmatic elements, pushing us perhaps in the direction of implicitures and away from generalized conversational implicatures—see Clifton and Ferreira, "Ambiguity in Context," *Language and Cognitive Processes*, IV (1989): 77–104.

[22] This expectation of connectivity will also inform our initial interpretations of the second and third occurrences of 'and' in (i). These expectations will remain in force and help give shape to the interpretation of the third 'and', but they will be set aside in favor of a straight truth-functional interpretation of the second.

tent. Therefore, even though reconception of the pragmatic content of (i) in terms of prepackaged associations helps us around the formal problems created by SGA, it leaves us with the familiar intuitive problems, a result that should embolden the proponent of TSA.

In fact, this proponent might well argue that rumors of TSA's death were greatly exaggerated. TSA remains, albeit in modified form. The argument in its new form will be identical with the old through (P3), but (P4) will be replaced by an assertion of OLPA. OLPA and (P2) generate a semantic content that cannot be true when the conjuncts in the antecedent of the conditionals are both true, undermining (P1). Therefore, we are forced to give up either (P1), (P2), or OLPA. Of these, it appears that (P2) loses out once again, and with it, the Gricean compromise approach.

The Gricean has a principled response to the argument in this new form. Underpinning this form of TSA is the assumption that our intuitive reaction to the utterance is keyed to the semantic content, which if true would militate against Grice's account of this content. This assumption is false, however. If we are to do justice to the complexity of utterance interpretation, we must recognize that it proceeds at two levels. The initial interpretation keys to the words used and involves identification of their semantic content. As this proceeds, the interpreter associates terms with the prepackaged associations she expects them to carry in conformity with OLPA. As these pragmatic elements are associated with the utterance on the fly, they are treated as provisional and defeasible parts of the total content, and so are open to modification or cancellation. These processes run in parallel, with the output of pragmatic processes[23] figuring into the total content that emerges, and the output of semantic processes figuring as input into pragmatic processes and as parts of the total content. As the total content emerges, it constrains how we take the output of semantic processes: if this output conflicts with the total content, we might treat the semantic output as an input to a pragmatic process or we might allow it to force a modification in the total content. Our decision will be made quickly and will be open to subsequent revision.[24]

---

[23] These processes include what Bach calls "completion" and "expansion"—see Bach, "Conversational Implicature." Cf. Récanati's processes "saturation" and "strengthening"—see Récanati, "The Pragmatics of What Is Said," p. 102. See also Récanati, *Direct Reference*, pp. 233–66; and Sperber and Wilson, pp. 181ff.

[24] Parallel processing is another respect in which the model advocated in this essay agrees with on-line processing models. Constraint-based, on-line models generally assume parallelism. See M. Tanenhaus and J. Trueswell, "Sentence Comprehension," in Joanne L. Miller and Peter D. Eimas, eds , *Speech, Language, and Communication* (San Diego: Academic, 1995), pp. 217–62. It should be noted, however, that allowance for parallel treatment of semantics and pragmatics in utterance interpretation does not necessarily imply computation of a full representation at both levels; rather, it

Semantic and pragmatic processes support the overall interpretation and both are essential to it; typically, neither process supplies the whole story about the utterance. In general, semantic content guides us as we introduce pragmatic elements; as we construct our interpretation of the total content of the utterance, the total content becomes the sole focus of our attention. The semantic content is to the total content of an utterance as a bare wall is to a decorated surface—the wall can be exposed, accentuated with trim, or covered with paintings or posters, but it must be there if there is to be any sort of decorated surface. The wall is essential, but we aim at a decorated surface, and once we have that, we can concentrate on the decoration and ignore the unexposed wall. The same is true of the semantic content: we use it in constructing our interpretation of the utterance, and while some aspects of it may be prominent parts of the final product of the interpretive process, other parts may serve only as platforms for pragmatic content; semantic aspects of the latter type support our identification of the relevant pragmatic content and then give way to them. (Of course, if need be, we could get back to the semantic content, just as we can get back down to the bare wall.)

Especially important for our purposes, semantic and pragmatic processes remain distinct even while they contribute to the formation of a unified and coherent interpretation of the total content of the utterance. Further, it is the total content and not the semantic content that matters in the final analysis.[25] The intuitions that make (P1)

---

might be that the processing system accesses and uses both semantic and pragmatic aspects of the input in building a representation of the total content. For discussion of this idea in connection with syntax and semantics, see W. Marslen-Wilson, L. Tyler, and M. Seidenberg, "Sentence Processing and the Clause Boundary," in Levelt and Flores d'Arcais, eds., pp. 219–46. For additional discussion, see R. Clark and E. Gibson, "A Parallel Model for Adult Sentence Processing," in *Program of the Tenth Annual Conference of the Cognitive Science Society* (Hillsdale, NJ: Lawrence Erlbaum, 1988), pp. 270–78; J. Dinsmore, "A Model for Contextualizing Natural Language Discourse," in *Program of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 597–604; P. Gorrell, "Establishing the Loci of Serial and Parallel Effects in Syntactic Processing," *Journal of Psycholinguistic Research*, XVIII (1989): 61–73; Marslen-Wilson, "Functional Parallelism in Spoken Word Recognition," *Cognition*, XXV (1987): 71–102; M. St. Johns and J. McClelland, "Applying Contextual Constraints in Sentence Comprehension," in *Program of the Tenth Annual Conference of the Cognitive Science Society*, pp. 26–32; and M. Tanenhaus, G. Dell, and G. Carlson, "Context Effects and Lexical Processing: A Connectionist Approach to Modularity," in Jay L. Garfield, ed., *Modularity in Knowledge Representation and Natural Language Understanding* (Cambridge: MIT, 1987), pp. 83–108. Compare the discussion of interpretive levels in Récanati, *Direct Reference*, pp. 312–18, and propositional form and style in Sperber and Wilson, pp. 202–17.

[25] As such, this proposal closely resembles one described near the end of Récanati, "What Is Said," *Synthese*, CXXVII (2001): 75–91; however, I would argue that while this notion of semantic content need not be "entertained or represented" at any point in utterance interpretation, it is nevertheless available to the interpreter and

plausible in these cases are keyed to *total content*; interpretation of this is developed dynamically from semantic content, some of which gives way to pragmatic elements as the interpretation goes forward.[26] Thus, difficulty at the semantic level will not threaten (P1) so long as it is not represented at the level of total content, and in these cases it is not. Therefore, recognition of utterance interpretation as a bilevel parallel process enables us to embrace (P2) and OLPA without threatening (P1), thereby establishing that the second version of TSA is also unsound.

III. A GRICEAN REJOINDER: PART TWO

As we have just seen, (i) is not a counterexample to Grice's view of conversation. His view is flexible enough to join with an on-line and parallel model of utterance processing, and together they supply an independently principled account of the intuitively nonproblematic character of (i). With the threat posed by (i) and its ilk dispatched, TSA is left without teeth. But while the argument will do no more damage, the same is not true of the spirit that motivated it, and so to that we now turn.

We have established that Grice's account of conversation can accommodate complex data like (i), but we have not addressed a deeper question, namely, whether TSA fairly represents the Gricean project. As we noted, TSA depends on processing readings of certain key terms, for example, 'calculate' and 'generate'. These terms signal the relevance of a psychological critique to some, but I am not among this group. These terms are ambiguous between psychological and nonpsychological readings, and Grice employed them in their psychologically innocent senses. His account is a contribution to the *logic* of conversation and not to its psychology. TSA conceals an equivocation, and in what follows, I argue that for this reason it fails to touch on the main business of Grice's account of conversation.

By emphasizing logic in his account, Grice calls attention to the structural character of conversation that supports inference and understanding. This character is viewed as general, underpinning conversational episodes in which the cooperative principle is observed. As we have seen, part of the logical story involves the truth-functional treatment given to connectives, but this provides a somewhat mis-

---

so has psychological reality in this sense.

[26] One might complain at this point that I have not indicated how Grice would have this process of replacement go; however, for reasons to be given in the next section, I do not believe that Grice must supply this. This would be a psychological story, and that is not what his account is intended to provide.

leading clue about the general nature of the project. His general concern with inferential structures is not as abstract as this, but is rather more concrete and practical; in particular, it focuses on the practical rationality of human communication. As Grice remarks, one of his "avowed aims is to see talking as a special case or variety of purposive, indeed rational behavior" (*op. cit.*, p. 28). Elsewhere, he is even more direct: "it is the rationality or irrationality of conversational conduct which I have been concerned to track down" (*op. cit.*, p. 369). The logical structure of language is important because it plays a fundamental role in the overall structure of conversation, and it is this structure that underwrites inferences about meanings against the backdrop of practical rationality. The logic of conversation, then, is a practical logic with rationality and not truth as its dominant metric, although truth remains a crucial element because of the implications it has for the rationality of conversational participants.

More detail about this logic is in order, although it can be no more than a sketchy outline at best here. Let us begin by taking a logic to be an inference-modeling system comprising primitive elements and inference rules, and a conversation to involve a sequence of utterances. Utterances need not involve language for Grice, but let us simplify here and focus on those that involve sentences in some language *L*. An utterance of this sort can be understood as the tokening of a sentence of *L* in a context, or more simply, a sentence token/context pair, where the context specifies properties such as the space-time coordinates of the utterance event and the producer of the token. Together, this pair introduces a specific proposition into the conversation, namely, the semantic content of the utterance. On the assumption that the participants in the conversation understand *L* and hear the utterance, they will likely have an attitude toward the semantic content introduced by the tokening of the sentence in that context; for example, they might believe it, deny it, and so forth. In addition, other propositional attitudes will be central to the conversation. The speaker will have intentions execution of which give rise to this utterance, as well as beliefs about his utterance, the audience, the context, and what this utterance will convey to his audience in this context. The audience will also have intentions that guide their participation in this episode, as well as beliefs about the utterance, the context, the speaker, and what the speaker might have intended to convey with that utterance in this context. These articulated assemblies of propositional attitudes are all *centered* on the utterance, in that the propositions involved must include as a constituent either the utterance or some aspect of the episode that is made relevant by the utterance. We now have what we need to provide a rough model of conversation. Take a *conversation* to be a sequence of utterances each of which serves as the center for an articulated complex of attitudes that the conversational participants have toward propositions about the constituent sentence tokens, contexts, and other participants.

A conversation is thus a sequence of utterances and an articulated and temporally extended assembly of propositional attitudes centered on those utterances. These are the elements that figure into the conversational inferences modeled by the logical system. Turning to the inference rules, we find that in this system, the rules divide into two different but related types. First, there are those that distinguish arrangements of propositional content that are truth conducing from those that are not. These rules identify clusters where the intended meaning of an utterance—that is, its total content—follows via abduction from the other propositions centered on the utterance. Here the inference rules of nondeductive logic and first-order deductive logic will apply. In this system, though, there is the added dimension of the attitudes taken toward the propositions by conversational participants.[27] These attitudes guide the participants' behavior and figure as constituents in the contents of the other propositional attitudes. The second type of inference rule identifies those attitude assemblies where belief in the total content of the utterance is rational, given belief in the other centered propositions. From the perspective of conversational success, inferences concerning rationality are more significant, since rational belief in the total content is generally what the speaker intends to make possible with her utterance. These two rule types are, of course, closely related—if the relations among the propositional contents conform to the proposition-level inference rules, then belief in the intended meaning will likely be rational, given belief in the other propositions.[28] It is as inference rules of the second type that we must see the cooperative principle and its attendant maxims. The cooperative principle is an overarching rule that underwrites the rationality of certain beliefs about the participants and the utterances. The maxims also underwrite rationality of belief, given other beliefs about the utterances, contexts, and participants. In general, these two types of inference rules codify patterns of nondeductive inference that are exemplified in episodes of successful communication, that is, communicative episodes where the total content to be conveyed by the utterance is actually conveyed.

---

[27] For a logical system that is similar in this respect, see Roderick Chisholm, *Theory of Knowledge*, Third Edition (Englewood Cliffs, NJ. Prentice Hall, 1989).

[28] These beliefs would not suffice for rationality if, for example, the agent in question also had other beliefs that were inconsistent with them. This reveals the essential nondeductive character of the attitude level inference rules.

Grice's account, then, is a rational reconstruction of conversation designed to reveal its logic. It is in light of this that one should understand terms like 'calculate' and 'generate'. They do not signal psychological maneuvers; rather, they signal the presence of truth and rationality conducing structural relationships among conversational elements. Their use here is similar to their use in mathematics, where one might speak of a theorem *generated* by a set of axioms, or a numerical claim *calculated* from other claims. To the extent that these terms do apply to our behavior as conversational participants, they describe what the propositions make possible and not what we actually do in thinking about them. With this in mind, let us return to TSA. We note that it depends on folding implicatures into an ongoing interpretation to solve a problem at one point and then assessing the connectives at a subsequent point. But this way of speaking betrays the fact that it construes the various propositions involved here, namely, the semantic and pragmatic contents associated with the embedded and embedding sentences, as temporally related in an extended processing event. But this is not psychology, it is logic, and so the standard to apply is rationality and not psychological plausibility. From this perspective, the semantic content of (i) conforms to (P2), warts and all, but an utterance of it centers a cluster of propositional attitudes that, taken together, provide abductive support for rational belief in the total content. There is no semantic problem posed by (i) that threatens the rationality of belief, given the other propositions centered on it.

"But surely," one might say, "it is in support of rational belief that the logical rubber meets the psychological road. TSA establishes that there is no way to realize a conversation involving utterance of (i) psychologically and still respect Grice's central tenets. His account is incompatible with empirically acceptable processing models and so must be set aside." There are two problems with this view. First, as I argued in the previous section, Grice can help himself to an independently justified processing model that can account for our intuitive feelings about (i) within the constraints supplied by his logic. This points to the source of the second problem, namely, the nature of those constraints. The critic would have us see these as rigid, forcing the processing model to be very much like and perhaps even identical with the logical model he describes. But identity is too strong. The exact relationship between the logic and psychology of conversation will likely be complex, but as we noted earlier, the logic must at least be compatible with the psychology, that is, the story about the inferential structure of conversation must be compatible with the cognitive story about processing if the logical story is to be sustainable. However complex the relationship turns out to be, though, it seems unlikely that the specific details of the processing model will be forced by the logic. Of course, the processing model *could* be a close implementation of the Gricean "calculations" that trace logical relations among propositional attitudes, but they need not be, and evidence suggests that this is unlikely. I believe that the correct processing story will be more akin to the one I have told, supplemented perhaps by accessibility considerations of the sort adduced in Récanati.[29] Whatever the cognitive story may turn out to be, the processes that figure into it are likely manifold and heterogeneous. Grice's goal was not to describe these processes, but only to plumb the logic that supported them; since there is no reason at this point to believe that his logical search should have also issued in descriptions of these processes, the key to acceptability of his account is compatibility. TSA sets out to prove incompatibility, but as we have seen, it fails.

Some support for the view that the relation between logic and psychology is loose can be gleaned from an examination of the rhetoric of argument. When one wishes to convey an argument to an audience, there will typically be many ways to present it. One must respect the argument, but one need not present it in its most logically revealing form. The logic of the argument underdetermines its expression and interpretation, constraining them without forcing them. This also seems true in the case of conversation, where its logic underdetermines its psychological realization. There are a number of possible processing models that could be the model that conversational participants instantiate, given the Gricean logic, including the one described in the preceding section. Further, it seems that this will be true regardless of the logical account that one endorses. Thus, there is reason to believe that compatibility and not identity should be our focus in relating the logic and psychology of conversation.

One might go further, however, and reject logic altogether as entirely irrelevant to conversation theory. This, however, would be a grave error. No defensible empirical account of conversation will be without an underlying logic. The elements of the conversational story relate to one another in ways that support inferences about meanings—this much is clear—and these patterns of inference depend on structural arrangements that ground the logic of conversation. Of course, there are many who dispute Grice's logic, and I have supplied little to assuage their concerns, especially given the limited focus on TSA. Indeed, the logic that underwrites the mature science of conversation may well be quite different from what Grice has limned, but TSA does not supply any reason to believe this.

---

[29] "The Alleged Priority of Literal Interpretation."

IV. CONCLUSION

Critics of Grice regularly conflate the logic of conversation with the psychology of conversation. Since his is an account of the pragmatics of language and so focuses on language use, they have taken psychology to be an appropriate lens for examining his view. This, however, is a mistake; Grice seeks to supply a rational reconstruction of the structure of conversational pragmatics and not a model of its psychological realization. This is *logic*, not psychology. That said, it is important to note that, while different, they are crucially related since the logic of conversation supplies the framework for the processing model. Within the framework, the conversational maxims serve as constraints on propositional attitude assemblies and so serve not as psychological laws but as logical laws that underwrite ascriptions of rationality. As its foundation, this framework employs a minimalist semantics for sentences, as reflected in (P2).

To the extent that TSA depends on a psychological reading of Grice's project, it is mistaken and misleading. One can interpret it, however, as an attempt to prove that Grice's logic is incompatible with any processing model that does justice to our intuitions. In particular, it purports to establish the incoherence of any attempt to assuage intuitive disturbances that utterances of (i) create by layering pragmatic content over a truth-functional core. Even charitably interpreted in this form, the argument fails. By combining Grice's story with a model of utterance processing that is on-line and parallel, we can demonstrate that TSA is unsound and readily accommodate our intuitions about examples like (i). This move also enables us to join the virtues of Gricean semantics as a platform for understanding the sentence's contribution to communication with the virtues of on-line accounts of utterance interpretation, a move that derives independent motivation from the empirical successes enjoyed by the on-line model. If we take our intuitions to give us insight into the total content of the utterance and we take seriously the on-line processes that yield our representation of this content, we can gratify our intuitions and maintain a commitment to minimalist semantics at the same time.

MICHAEL O'ROURKE

University of Idaho

COMMENTS AND CRITICISM

SAVOIR FAIRE*

In their article "Knowing How,"*[1] Jason Stanley and Timothy Williamson seek to argue, against Gilbert Ryle and others, that knowing how to do something is a special case of knowing that such-and-such is the case. On their view, the attribution

(1) Hannah knows how to ride a bicycle.

is true "if and only if there is some contextually relevant way *w* such that Hannah stands in the knowledge-that relation to the Russellian proposition that *w* is a way for Hannah to ride a bicycle" (430). In this note, I shall not challenge Stanley and Williamson's conclusion. I wish to explain, however, why their argument for it is quite unconvincing.

The argument starts from the linguistic claim that attributions of knowledge-how such as (1) "contain what are called embedded questions" (418). The attribution is thereby assimilated to a family whose most closely related members include

(2) (a) Hannah knows whom to call for help in a fire.
(b) Hannah knows why to vote for Gore (418)

In these last examples, the verb 'to know' precedes an interrogative particle, which in turn precedes an infinite clause. Stanley and Williamson then appeal to a general semantic theory of sentences involving embedded questions, according to which a sentence such as (2a) will be true if and only if Hannah knows some proposition that could (in the relevant context of utterance) answer the embedded question. She will know whom to call for help in a fire if and only if she knows a proposition that answers the question 'Whom to call for help in a fire?' or (more idiomatically) 'Whom should I call for help in a fire?' Similarly, she knows why to vote for Gore if and only if she knows a proposition that answers the question 'Why vote for Gore?'[2] As Stanley

---

*I am very grateful for comments from Lucy Baines, Jonathan Barnes, Corine Besson, Stephen Everson, David Glot, Michael Nicholson, Jason Stanley, Timothy Williamson, and, especially, David Wiggins.

[1] This JOURNAL, XCVIII, 8 (August 2001): 411–44. Unadorned page references will be to this article.

[2] See Stanley and Williamson's discussion of the "mention-some" readings of embedded questions (425–26). Reflection on example (2b) reveals, though, a complexity in the relevant notion of knowing a proposition that answers a question. For what

0022-362X/03/0003/158–66

and Williamson observe, on this treatment of embedded questions, the sort of knowledge ascribed in the examples in (2) is propositional knowledge. The sentence (2a) differs from 'Hannah knows that she should call her parents for help in a fire', not in ascribing some nonpropositional variety of knowledge, but in failing expressly to identify the proposition that she is said to know. Stanley and Williamson proceed to castigate philosophers who have followed Ryle in supposing that knowledge-how is not a species of knowledge-that for failing to acquaint themselves with current treatments of the semantics of the English expression 'knowing how'.

The topic that interested Ryle, however, was the nature of the knowledge that is ascribed when we say that Hannah knows how to ride a bicycle, not the particularities of its ascription in English. Stanley and Williamson are aware of this. As well as proposing a semantic theory for English sentences such as (1), they "present and defend [a] positive account of knowledge-how, according to which it is a species of knowledge-that" (412). It is this thesis—a metaphysical thesis, as one might call it, about the nature of knowledge-how—that is the real locus of disagreement with Ryle, who had cited certain instances of knowledge-how as counterexamples to the "intellectualist legend" that all knowledge is propositional.[3] Stanley and Williamson understand their metaphysical thesis to be incompatible with any claim to the effect that a person's knowing how to do something is a matter of his being appropriately related to an activity. This emerges most clearly from their discussion of David Carr, to whom the latter

---

is it to know a proposition that answers the question, 'Why vote for Gore?'? A true answer to this question may be, 'Because he will act to protect the environment'. But while Hannah may be said to know to vote for Gore because he will protect the environment, this knowledge is not obviously propositional. It is the knowledge that one has when one knows to do something for a reason. Of course, the knowledge that Gore will act to protect the environment is straightforwardly propositional. Yet Hannah might know that Gore will act to protect the environment without knowing that Presidents of the United States are elected at all. Perhaps she believes that they are appointed by the U.S Supreme Court; perhaps, indeed, she lacks even the concept of voting for someone. If that is her situation, however, it is surely not true that she knows why to vote for Gore This suggests that, in order to qualify as knowing why to φ, a subject needs, not merely to know the propositional content of a true answer to the question 'Why φ?', but also to know that that proposition is a true answer to the question Since, however, this additional knowledge is still knowledge-that, we need not pursue this matter further here.

[3] See Ryle, *The Concept of Mind* (Chicago: University Press, 1949), chapter II. "When we speak of the intellect or, better, of the intellectual powers and performances of persons, we are referring primarily to that special class of operations which constitute theorising. The goal of these operations is the knowledge of true propositions or facts.... The main goal of this chapter is to show that there are many activities which directly display qualities of mind, yet are neither themselves intellectual operations nor yet effects of intellectual operations" (p. 26).

metaphysical claim is attributed (416–17). They advertise section II of their paper as "undermining" any theory like Carr's. The section in question is devoted to defending their own positive account of knowledge-how, and contains no independent attack on Carr's view. The advertisement, then, is out of place unless their metaphysical account of knowledge-how is incompatible with Carr's, so that any support for their view *ipso facto* undermines his. One can, though, readily understand their thesis in such a way that it is incompatible with Carr's, and I shall assume that it is so understood.

What evidence do Stanley and Williamson present in support of their metaphysical thesis? The evidence seems to consist in the existence of sentences each of which (a) is used to attribute knowledge-how, and (b) is such that the best semantic theory for the language to which it belongs will construe its knowledge-verb as expressing (in that sentential context) a relation between a person and a proposition. (They cite example (1) above as having these attributes.) I shall not challenge the methodological claim that the existence of sentences meeting conditions (a) and (b) provides support for Stanley and Williamson's metaphysical thesis. Three observations, however, are in order. First, by whatever token their metaphysical thesis is supported by the existence of sentences meeting conditions (a) and (b), Carr's thesis will be supported by the existence of sentences meeting (a) and a new condition (c), which reads as follows: being such that the best semantic theory for the language to which it belongs will construe its knowledge-verb as expressing (in that sentential context) a relation between a person and an activity. Second, since Stanley and Williamson's metaphysical thesis and Carr's thesis are contraries, any evidence that supports the one undermines the other. In evaluating Stanley and Williamson's thesis, then, it is not enough to investigate whether there are indeed examples meeting conditions (a) and (b); one must also investigate whether there are examples meeting conditions (a) and (c). Third, the quest for such examples cannot be confined to English sentences. This is because the metaphysical thesis concerns the nature of knowledge-how, not the semantics of 'knows how'.

As soon as one embarks on this quest, one must notice that English is far from typical in employing an interrogative particle ('how') to ascribe the knowledge that one has when one knows how to ride a bicycle, or to swim, or to speak a language. A comparison with French is instructive. The French verb for the relevant kind of knowledge is *savoir*, and the *savoir* + interrogative word + infinitive construction is common in French. One can say, for example,

(3) Jean sait quoi manger.
Jean knows what to eat.

(4) Pierre sait comment traverser le fleuve en nageant.
Pierre knows how to swim across the river.

The analysis that Stanley and Williamson recommend for sentences containing embedded questions works well in these cases. It is plausible to hold that (3) is true if and only if Jean knows a proposition which qualifies (in the context of utterance) as answering the question 'Quoi manger?' (a perfectly standard direct question in French). Similarly, (4) is true if and only if there is a question of how to swim across the river (a problem to be solved, as one might say), and Pierre knows something which answers this question (that is, which solves the problem). So (4) might be true if, for example, Pierre knows that the way to swim across the river is to swim to the red buoy and then head straight for the northern bank. This knowledge is indeed propositional. When a French speaker ascribes the sort of knowledge-how in which Ryle was interested,[4] however, he does not use an interrogative construction at all, but instead employs a bare infinitive:

(5) Elle sait monter à vélo.
She knows how to ride a bicycle.

(6) Il sait nager.
He knows how to swim.

(7) Il sait parler français.
He knows how to speak French.

Since the interrogative word *comment* is not to be found in them, these French attributions afford no support to Stanley and Williamson's hypothesis that the knowledge being attributed when we say that someone knows how to ride a bicycle is knowledge of a proposition which could answer a 'How to?' question.

There is a corresponding difference among the French constructions that can translate 'learn how', and so forth. One can say,

(8) Il a appris comment traverser le fleuve en nageant.
He has learned how to swim across the river.

using the interrogative word + infinitive construction. And one would say this of someone who had learned to swim to the red buoy and then head north. In saying, however, that someone is simply learning how to swim, a French speaker will eschew the interrogative particle and use a preposition plus infinitive:

---

[4] His examples are knowing how to: play an instrument; prune trees; tie a reef-knot; cook a meal, swim; play chess; crack jokes; reason; fish; prevail in battle. See Ryle, pp. 27–32.

(9) Il apprend à nager.
He is learning (how) to swim.

The fact that French does not provide examples of sentences which attribute the relevant kind of knowledge-how but which involve embedded questions is not by itself of great moment, although it shows that Stanley and Williamson would have difficulty in presenting their argument to a monoglot French speaker. The interest of (5) to (7) is, rather, that they are plausible examples of sentences which meet conditions (a) and (c), and which thereby constitute evidence against Stanley and Williamson's metaphysical thesis. For French sentences like (5) to (7) are most closely related, not to sentences like (3) and (4), but to other sentences involving a bare infinitive, such as

(10) Il peut nager.
He can swim.

(11) Nager, c'est merveilleux.
Swimming is marvellous.

In these cases, the semantic function of the infinitive is surely to refer to an activity. But if we treat the infinitives in (5) to (7) in the same way, then the finite form of *savoir* that they contain must be treated as expressing a relation between a person and an activity, so that these sentences will meet conditions (a) and (c).

The difference between (4) and (5) to (7) also bears on a possible defense of Stanley and Williamson's theory which postulates an unpronounced interrogative particle in the "deep structure" of (5) to (7). The fact that

(12) Il sait comment nager.

is a well-formed French sentence may encourage this defensive maneuver. Less convenient for this defense is the fact that (12) would mean either something like 'He has solved the problem of getting to swim'—compare (4)—or, perhaps, 'He knows in what particular manner to swim'—compare 'He knows how to dress'. While the knowledge attributed under each of these interpretations may well be propositional, in neither case is it necessary or sufficient for knowing how to swim. The defensive maneuver would also have to explain why

(13) *Il apprend comment nager.

is ill formed.

French is not alone in eschewing interrogative words in favor of bare infinitives when ascribing the knowledge that people have when they know how to swim, to ride bicycles, to speak languages, and so forth. In Greek we find

(14) οὔπω σωφρονεῖν ἐπίστασαι;
Do you not yet know how to exercise self-control? (Aeschylus, *Prometheus* 982).

In Latin we have

(15) Vincere scis, Hannibal.
You know how to conquer, Hannibal (Livy 22, 51, 4).

In neither of these cases is there any sign of an interrogative particle; so there is no temptation to interpret these sentences as attributing knowledge of a proposition that can serve as answer to a question or a problem. Moreover, attributions of knowledge-how in these languages appear to belong with constructions in which a bare infinitive is used to refer to an activity, as in Plautus's warning:

(16) Incipere multo est quam impetrare facilius.
Beginning is much easier than getting to the end (*Poenulus* 5, 2, 14).

Stanley and Williamson may try to defuse the threat that these cases pose for their metaphysical thesis by arguing that the infinitives that occur in (5) to (7), (14) and (15) do not, despite appearances, refer to an activity. They may, indeed, seek an alternative semantic description under which these cases meet condition (b). (Finding such a description seems to be the only way of excluding the possibility that the sentences' optimal semantic description will support a metaphysical thesis which is just as incompatible with their preferred account of knowledge-how as Carr's account is.) The quest for an analysis under which these cases meet condition (b) is far from hopeless: a sentence need not contain an embedded question to attribute propositional knowledge.[5] What sustains it, though, is the fact that each of the verbs in (5) to (7), (14), and (15) *can* be used to ascribe propositional knowledge. We see here the importance of Stanley and Williamson's conjecture that "the uses of 'know' in ['Hannah knows how to ride a bicycle'] and ['Hannah knows that penguins waddle'] are translated by the same word" in all natural languages (437).

Russian, however, provides a counterexample to this conjecture. The verb used to attribute propositional knowledge in Russian is знать (*znat'*), and this verb is also used to ascribe knowledge of the answers to questions. Thus we have:

---

[5] In discussion, Stanley has suggested that the bare infinitive in, for example, (5) does not (as one might at first expect) serve to designate a kind of activity, but instead signifies a range of propositions over which the verb *savoir* quantifies.

(17) Я знаю, что я должен плавать.
(Ya znayu, chto ya dolzhen plavat')
I know that I must swim.

(18) Он знает, зачем избрать Путина.
(On znaet, zachem izbrat' Putina)
He knows why to vote for Putin.

(compare (2b) above). In attributing knowledge-how, though, one uses the unrelated verb уметь (*umet'*), which has a quite different grammatical distribution. It can be complemented neither by a что-clause (that is, a 'that'-clause), nor by an indirect question, but must govern an infinitive. Thus we have

(19) Я умею плавать.
(Ya umeyu plavat')
I know how to swim.

Like 'knows how', the verb уметь does not attribute occurrent ability. One could say, for example, of a broken-armed man

(20) Он умеет плавать, но он не может плавать.
(On umeet plavat', no on ne mozhet plavat')
He knows how to swim, but he is unable to swim.

Indeed, уметь is clearly an intellectual verb. The cognate noun ум (*um*) means 'mind' or 'intellect', and the cognate adjective умный (*umniy*) means 'clever', 'wise', or 'intelligent'. A similar difference is found between verbs that translate 'to learn'. Quite different verbs are found in

(21) Он научился плавать.
(On nauchilsya plavat')
He learned how to swim.

and

(22) Он узнал, как Троцкий умер.
(On uznal, kak Trotskiy umer)
He learned how Trotsky died.

As with the verbs of knowledge, the verb in (21), научиться (*nauchit'sya*) must govern an infinitive while the verb in (22), узнать (*uznat'*), is complemented by a 'that'-clause or an indirect question. For this reason, there seems to be no prospect at all that the best semantic theory for Russian will construe the verbs of knowledge and learning in (19) to (21) as expressing propositional knowledge or propositional learning, and every prospect that it will construe them as expressing relations between a person and an activity.

The linguistic evidence bearing on Stanley and Williamson's thesis is, then, equivocal. Some evidence may be found which supports it; some evidence may be found which undermines it. This already shows that a *quietus* for the contrary theses of Ryle and Carr would be premature, but it leaves us in a quandary. What are we to think about the nature of knowledge-how?

The linguistic evidence bearing on this question deserves attention from those more expert than me; this is an area where a trawl through cross-linguistic data might be of real philosophical interest. But let me conclude by tentatively mentioning the merits of an answer that is liable to be overlooked. Semantic theorists are loath to postulate ambiguities, and their reluctance to do so is in general healthy. When different instances of an apparently univocal construction translate very differently, however, the presumption of univocity may be questioned. In the present case, we have found that the 'knows how to' construction in English covers attributions which some other languages mark very differently. On the one side, we have the sort of case which (4) exemplifies—the *savoir comment faire* cases, as we may call them—in which a person is said to know the solution to a problem. On the other side, we have the sort of case exemplified by (5) to (7), (14), (15), (19), and (20)—the *savoir faire* cases. The knowledge attributed in *savoir comment faire* cases may be a relation to a proposition that can serve as an answer to the component 'How to?' question, just as Stanley and Williamson propose. For all that, however, the knowledge attributed in *savoir faire* cases may be a relation between a person and the activity indicated by the component infinitive.

What evidence might there be for this apparently ungainly view? One point in its favor is the ease with which English speakers learning French come to translate the different sorts of case correctly. They seem to possess a latent apprehension of the two sorts of case that acquisition of the new language enables them to mark. Another point is suggested by Stanley and Williamson's own appeal to "the possibility of conjunction" to show that the objects of propositional knowledge and of knowledge-how belong to the same "semantic type" (430):

(23) John knows both how to twitch his ears and that his mother is sickened by facial tricks.

I am not alone in finding (23) zeugmatic, like (24), while (25) is not:

(24) She arrived in floods of tears and a Bath chair.

(25) John knows both that Trotsky was murdered and how he was killed.

This difference is just what one would expect if 'knows how he was killed' adverts to propositional knowledge while 'knows how to twitch his ears' does not. Third, there is the odd way we use the verbal noun 'know-how'. We say

(26) Knowing how to fix a leak is part of any competent plumber's know-how.

but we do not say

(27) Knowing how Trotsky was killed is part of any competent historian's know-how.

just as we do not say

(28) Knowing why Trotsky was killed is part of any competent historian's know-why.

The cases where we use 'know-how' seem to correspond quite closely with those that might be translated using *savoir faire*.

The restricted thesis that *some* ascriptions in the form '*x* knows how to *F*' do not ascribe propositional knowledge would continue to serve Ryle's ulterior philosophical purposes. He wanted to dispel "the intellectualist legend"—the idea that instances of knowledge are always instances of knowing that such-and-such is the case. In dispelling the legend, it mattered that he could identify a class of instances of knowledge that are not cases of knowing a proposition. It also mattered, if the legend was not to survive in some emended or qualified form, that the purported counterexamples should be reasonably common and central. But it never mattered that those counterexamples should include all of those attributed using the 'knows how to' construction in English. It is, indeed, striking that all the examples he gives of the sort of knowledge-how that he has in mind would naturally be translated into French using a *savoir faire* construction.[6]

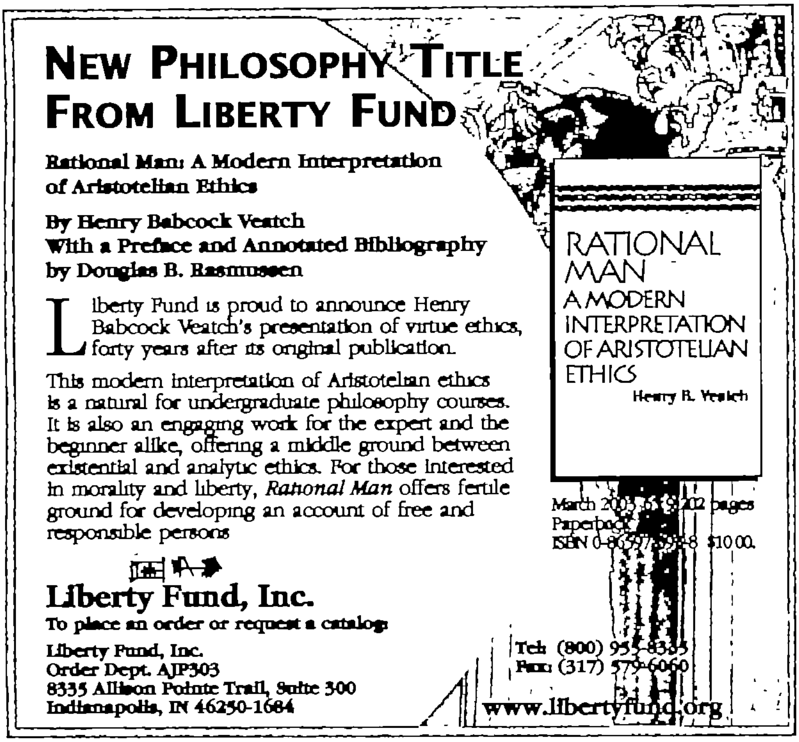
As I say, the whole matter needs further investigation, and not just by linguists. Perhaps the issue can be decided by more "philosophical" considerations, akin to the regress argument that Stanley and Williamson rightly reject at the start of their paper. But Ryle's hypothesis that knowledge-how is not a species of knowledge-that is not refuted by Stanley and Williamson's article.

IAN RUMFITT

University College, Oxford

---

[6] See again the examples listed in footnote 4.



NEW PHILOSOPHY TITLE FROM LIBERTY FUND

Rational Man: A Modern Interpretation of Aristotelian Ethics

By Henry Babcock Veatch
With a Preface and Annotated Bibliography by Douglas B. Rasmussen

Liberty Fund is proud to announce Henry Babcock Veatch's presentation of virtue ethics, forty years after its original publication.

This modern interpretation of Aristotelian ethics is a natural for undergraduate philosophy courses. It is also an engaging work for the expert and the beginner alike, offering a middle ground between existential and analytic ethics. For those interested in morality and liberty, *Rational Man* offers fertile ground for developing an account of free and responsible persons.

March 2003. 5 [illegible] 9. 202 pages
Paperback
ISBN 0-86597-[illegible]-8 $10.00.

Liberty Fund, Inc.

To place an order or request a catalog:

Liberty Fund, Inc.
Order Dept. AJP303
8335 Allison Pointe Trail, Suite 300
Indianapolis, IN 46250-1684

Tel: (800) 955-8335
Fax: (317) 579-6060
www.libertyfund.org

THE JOURNAL OF PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org*.

* * *

PURPOSE & HISTORY	RECENT & FORTHCOMING ISSUES
CONTACTS AND ADDRESSES	ELECTRONIC ISSUES
BOARD OF TRUSTEES	INDEX 1904–PRESENT
EDITORIAL BOARD	RIGHTS AND PERMISSIONS
MEMOS TO AUTHORS	LECTURES & CONFERENCES
SUBSCRIPTIONS & ADVERTISING	PHILOSOPHY LINKS

* * *

Classic Works of Philosophy

New from Thoemmes Press

Jeremy Bentham's Auto-Icon and Related Writings

Edited by JAMES CRIMMINS

At University College, London, the corpse of Jeremy Bentham can still be seen seated inside a glass cabinet. This was in accordance with his own wishes, but scholars have never had the material available to understand his reasons. The heart of this volume, however, is Bentham's never-before-published justification for the auto-icon, as well as the medical, philosophical, and religious contexts for his thoughts.

202 pages, 1-85506-[illegible]-1
Cloth $[illegible].00

Social Rights and Duties

LESLIE STEPHEN

These volumes collect twelve lectures, originally delivered between 1890 and 1895, in which Stephen applies his evolutionary ethics to a series of social and political issues, creating an important historical record of England's early ethical movement.

[illegible] pages, 2 Volumes, 1-85506-[illegible]-7
Cloth $225.00

Dictionary of Nineteenth-Century British Philosophers

Edited by W. J. MANDER and ALAN P. F. SELL

Containing more than 600 biographical entries on figures of the nineteenth century, this important new work provides the most detailed and comprehensive reference work on Victorian thought and thinkers ever published.

1000 pages, 2 Volumes, 1-85506-[illegible]-5
Cloth $[illegible].00

Descartes in Seventeenth-Century England

RENÉ DESCARTES

Edited by Daniel Garber and Roger Ariew

These volumes contain Descartes's main works in their first English translations, as well as critiques of his philosophy both translated into English and English in origin. Other works in the set bring together works by Cartesians in English translation, works by English thinkers influences by Descartes, and the standard seventeenth-century Descartes biographies in their English translations.

[illegible] pages, 10 Volumes, 1-85506-[illegible]-3
Cloth $1275.00

Distributed by

The University of Chicago Press

1427 East 60th Street, Chicago, IL 60637 www.press.uchicago.edu

Classic Works of Philosophy

New from Thoemmes Press

Samuel Johnson

His Career and Writings

Edited by HERBERT SCHNEIDER and CAROL SCHNEIDER

Arguably one of the three main thinkers of pre-independence America, Samuel Johnson was the chief exponent of Berkeley's idealism and the first president of the religious college that later became Columbia University. Containing his autobiography and correspondence with other important figures of the period, his philosophical and clerical works, and a history of Columbia University, these volumes also contain a bibliography and index.

2217 pages, 4 Volumes, 1-84371-003-X
Cloth $485.00

Oeuvres Philosophiques d'Arnauld

ANTOINE ARNAULD

Edited by Elmar J. Kremer and Denis Moreau

This collection presents the specifically philosophical portions of Antoine Arnauld's work in the original French and will be of great interest to both historians of ideas and to present-day philosophers who find his views on intentionality and perception relevant to current debates.

2934 pages, 6 Volumes, 1-85506-967-9
Cloth $695.00

The Early American Reception of German Idealism

Edited by JAMES A. GOOD

It is often supposed that mid-nineteenth century American philosophy was provincial and unexciting—that it consisted of sleepy professors regurgitating the Common Sense Realism learned from Scotland. This set shows that there was a sophisticated coterie of American writers who were very well aware of the German Idealist revolution and philosophy, and sought to transplant it to American soil.

2628 pages, 5 Volumes, 1-85506-982-X
Cloth $635.00

An Examination of Plato's Doctrines

I. M. CROMBIE

This wide-ranging, detailed, and sophisticated work of Platonic commentary contains a comprehensive introduction to Plato's moral and political thinking, his philosophies of mind and religion, and his thoughts on the more technical topics of philosophy.

982 pages, 2 Volumes, 1-85506-986-5
Cloth $225.00

Distributed by

The University of Chicago Press

1427 East 60th Street, Chicago, IL 60637 www.press.uchicago.edu

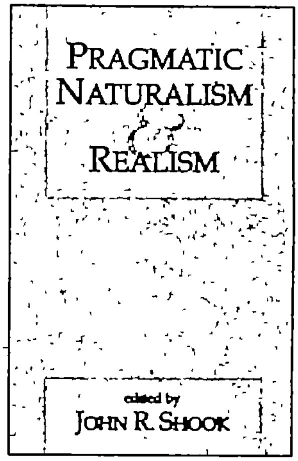
New Philosophy Titles from . . .

Prometheus Books

PRAGMATIC NATURALISM & REALISM
Edited by John R. Shook
325 pp • ISBN 1-57392-997-2 • HC $32

THE SCIENCE WARS
Debating Scientific Knowledge and Technology
Edited by Keith Parsons
300 pp • ISBN 1-57392-994-8 • PB $21

CONTROLLING TECHNOLOGY
Contemporary Issues
Edited by Eric Katz, Andrew Light, and William Thompson
550 pp (Illustrations)
ISBN 1-57392-983-2 • PB $34



PRAGMATIC NATURALISM & REALISM
edited by
JOHN R. SHOOK

. . . and

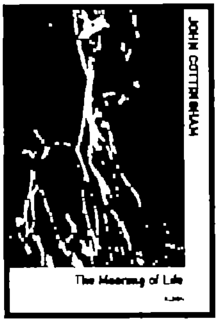
HB Humanity Books

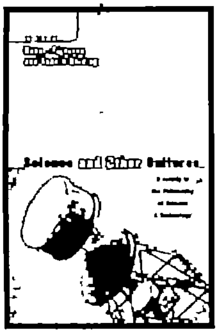
PASSION AND VIRTUE IN DESCARTES
Edited by Byron Williston and André Gombay
285 pp • ISBN 1-59102-006-0 • HC $59

KARL JASPERS ON PHILOSOPHY OF HISTORY AND HISTORY OF PHILOSOPHY
Edited by Joseph W. Koterski, S.J., and Raymond J. Langley
300 pp • ISBN 1-59102-002-6 • HC $65

MIND, BODY, AND FREEDOM
Patrick T. Mackenzie
140 pp • ISBN 1-59102-001-8 • HC $45

For exam copies contact editorial@prometheusbooks.com
Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com





Philosophy from Routledge

EPISTEMOLOGY
A Contemporary Introduction to the Theory of Knowledge
Second Edition
Robert Audi

"*Epistemology is simply the best textbook in epistemology that I know of.*"
—Thomas Vinci, Dalhousie University, Canada

368 PP • $22.95

ON THE MEANING OF LIFE
John Cottingham

The question "What is the meaning of life?" is one of the most fascinating, oldest and most difficult questions human beings have ever posed themselves. In this acute and thoughtful book, John Cottingham asks why the question vexes us so much and assesses some of the most influential attempts to explain it.

144 PP • $12.95

ACTION, EMOTION AND WILL
Revised Edition
Anthony Kenny

Action, Emotion and Will was first published in 1963, when it was one of the first books to provoke serious interest in the emotions and philosophy of human action. Almost forty years on, this is still essential reading for anyone interested in these topics.

192 PP • $24.95

THE LIFE OF THE MIND
An Essay of Phenomenological Externalism
Gregory McCulloch

"*...distinctive and original—non-cartesian, non-materialist, deeply humane—outlook on the mind...characteristic verve, clarity and good humour.*"
—Crispin Wright, University of St Andrews

176 PP • $25.95

SCIENCE AND OTHER CULTURES
Diversity in the Philosophy of Science and Technology
Edited by
Sandra Harding and Robert Figueroa

In this pioneering new book, original essays by leading philosophers explore an extensive range of diversity issues for the philosophy of science and technology.

320PP • $24.95

DEBATES IN CONTEMPORARY POLITICAL PHILOSOPHY
An Anthology
Edited by
Derek Matravers and Jonathan Pike

This book provides a comprehensive collection of influential essays that present a balanced survey of the major ideas that have come out of this area of study in the last two decades.

464 PP • $29.95

Routledge

1.800.634.7064 WWW.ROUTLEDGE-NY.COM

Outstanding Scholarship from Cambridge

Aquinas: Political Writings
R.W. Dyson, Editor
Cambridge Texts in the History of Political Thought
0-521-37569-X Hardback $60.00
0-521-37595-9 Paperback $22.00

Natural Rights and the Right to Choose
Hadley Arkes
0-521-81218-6 Hardback $28.00

Augustine: On the Trinity
Gareth Matthews, Editor
Cambridge Texts in the History of Philosophy
0-521-79231-2 Hardback $55.00
0-521-79665-2 Paperback $20.00

Classic and Romantic German Aesthetics
J.M. Bernstein, Editor
Cambridge Texts in the History of Philosophy
0-521-80639-9 Hardback $60.00
0-521-00111-0 Paperback $22.00

Nietzsche: Writings from the Late Notebooks
Rüdiger Bittner, Editor
Cambridge Texts in the History of Philosophy
0-521-80405-1 Hardback $50.00
0-521-00887-5 Paperback $18.00

The Cambridge Companion to Simone de Beauvoir
Claudia Card, Editor
Cambridge Companions to Philosophy
0-521-79096-4 Hardback $60.00
0-521-79429-3 Paperback $22.00

Radical Interpretation in Religion
Nancy K. Frankenberry, Editor
0-521-81686-6 Hardback $65.00
0-521-01705-X Paperback $24.00

Realism and Christian Faith
God, Grammar, and Meaning
Andrew Moore
0-521-81109-0 Hardback $65.00
0-521-52415-6 Paperback $23.00

More: Utopia
Second Edition
George M. Logan and Robert M. Adams, Editors
Cambridge Texts in the History of Political Thought
0-521-81925-3 Hardback $30.00
0-521-52540-3 Paperback $10.00

Visions of Politics
Quentin Skinner
Three-Volume Set
0-521-81382-4 Hardback $180.00
0-521-89075-6 Paperback $65.00

The Philosophy of Social Practices
A Collective Acceptance View
Raimo Tuomela
0-521-81860-5 Hardback $60.00

Vico: The First New Science
Leon Pompa, Editor
Cambridge Texts in the History of Political Thought
0-521-38290-4 Hardback $65.00
0-521-38726-4 Paperback $23.00

The Cambridge Companion to Duns Scotus
Thomas Williams, Editor
Cambridge Companions to Philosophy
0-521-63205-6 Hardback $65.00
0-521-63563-2 Paperback $23.00

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

800-872-7423
us.cambridge.org/philosophy

THE JOURNAL OF
PHILOSOPHY

Calcutta University Library
CAMPUS

VOLUME C, NUMBER 4
APRIL 2003

page

167 *Truth as Convenient Friction* Huw Price
191 *A Spectral Reflectance Doth Not a Color Make* C. L. Hardin
203 *Against Musical Ontology* Aaron Ridley

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX: (212) 932-3721

You may also visit our website at www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 4, APRIL 2003

12/8/03

TRUTH AS CONVENIENT FRICTION*

In a recent paper, Richard Rorty[1] begins by telling us why pragmatists such as himself are inclined to identify truth with justification:

> Pragmatists think that if something makes no difference to practice, it should make no difference to philosophy. This conviction makes them suspicious of the distinction between justification and truth, for that distinction makes no difference to my decisions about what to do (*ibid.*, p. 19).

Rorty goes on to discuss the claim, defended by Crispin Wright, that truth is a normative constraint on assertion. He argues that this claim runs foul of this principle of no difference without a practical difference:

> The need to justify our beliefs and desires to ourselves and to our fellow agents subjects us to norms, and obedience to these norms produces a behavioral pattern that we must detect in others before confidently attributing beliefs to them. But there seems no occasion to look for obedience to an *additional* norm—the commandment to seek the truth. For—to return to the pragmatist doubt with which I began—obedience to that commandment will produce no behavior not produced by the need to offer justification (*ibid*, p. 26).

Again, then, Rorty appeals to the claim that a commitment to a norm

---

*The first version of this paper was written for a conference in honor of Richard Rorty at the Australian National University in 1999. I am grateful to participants at that conference and to many subsequent audiences for much insightful discussion of these ideas.

[1] "Is Truth a Goal of Inquiry? Donald Davidson versus Crispin Wright," in his *Truth and Progress: Philosophical Papers*, Volume 3 (New York: Cambridge, 1998), pp 19–42.

0022-362X/03/0004/167–90

of truth rather than a norm of justification makes no behavioral difference.

This is an empirical claim, testable in principle by comparing the behavior of a community of realists (in Rorty's sense) to that of a community of pragmatists. In my view, the experiment would show that the claim is unjustified, indeed, false. I think that there is an important and widespread behavioral pattern that depends on the fact that speakers do take themselves to be subject to such an additional norm. Moreover, it is a behavioral pattern so central to what we presently regard as a worthwhile human life that no reasonable person would knowingly condone the experiment. Ironically, it is also a pattern that Rorty of all people cannot afford to dismiss as a pathological and dispensable by-product of bad philosophy. For it is conversation itself, or at any rate a central and indispensable part of conversation as we know it—roughly, interpersonal dialogue about "factual" matters.[2]

In other words, I want to maintain that in order to account for a core part of ordinary conversational practice, we must allow that speakers take themselves and their fellows to be governed by a norm stronger than that of justification. Not only is this a norm which speakers acknowledge they may fail to meet, even if their claims are well justified—this much is true of what Rorty calls the *cautionary* use of truth[3]—but also, more significantly, it is a norm which speakers immediately assume to be breached by someone with whom they disagree, *independently of any diagnosis of the source of the disagreement.* Indeed, this is the very essence of the norm of truth, in my view. It gives disagreement its immediate normative character, a character on which dialogue depends, and a character which no lesser norm could provide.

This fact about truth has been overlooked, I think, because the norm in question is *so* familiar, so much a given of ordinary linguistic practice, that it is very hard to see. Ordinarily we look through it, rather than at it. In order to make it visible, we need a sense of how things would be different without it. Hence, in part, my reason for beginning with Rorty. Although I disagree with Rorty about the behavioral consequences of a commitment to "a distinction between justification and truth," I think that the issue of the behavioral consequences of such a commitment embodies precisely the perspective we need, in order to bring into focus this fundamental aspect of the normative structure of dialogue.

---

[2] Irony aside, nothing here turns on whether by 'conversation' I mean the same as Rorty. For me, what matters is the role of truth in the kind of interpersonal linguistic interaction I will call factual or assertoric dialogue, or simply dialogue. I do not claim that dialogue exhausts conversation, in Rorty's sense or any other. I used scare quotes on 'factual' above in anticipation of the suggestion that the notion of factuality in play might depend on that of truth, in a way which would create problems for my own account of the role of truth in dialogue. There is no such difficulty, in my view. On the contrary, I take the perceived "factuality" or "truth-aptness" of the utterances in question to be part of the explanandum of the kind of account here proposed; cf. footnote 23.

[3] "Pragmatism, Davidson, and Truth," in *Objectivity, Relativism, and Truth: Philosophical Papers*, Volume 1 (New York: Cambridge, 1991), pp. 126–50, see p. 128.

In sharing Rorty's concern with the role of truth in linguistic practice, I share one key element of his pragmatism. But my kind of pragmatism about truth is not well marked on contemporary maps, and hence my second reason for beginning with Rorty. Rorty has explored the landscape of pragmatist approaches to truth more extensively than most pragmatist writers, past or present, and at different times has been inclined to settle in different parts of it. By locating my own kind of pragmatism with respect to views that Rorty has visited or canvassed, I hope to show that there is a promising position that he and others pragmatists have overlooked.

As noted, my view rests on the claim that a norm of truth plays an essential and little-recognized role in assertoric dialogue. In pursuit of this conclusion, it will turn out to be helpful to distinguish three norms, in order of increasing strength, roughly: sincerity, justification, and truth itself. Though somewhat crudely drawn, these distinctions will suffice to throw into relief the crucial role of the third norm in linguistic practice. My strategy will be to contrast assertion as we know it with some nonassertoric uses of language. In these latter cases, I will argue, the two weaker norms still apply. Moreover, it turns out that some of the basic functions of assertoric discourse could be fulfilled in an analogous way, by a practice which lacked the third norm. But it will be clear, I hope, that that practice would not support dialogue as we know it. What is missing—what the third norm provides—is the automatic and quite unconscious sense of engagement in common purpose that distinguishes assertoric dialogue from a mere roll call of individual opinion. Truth is the grit that makes our individual opinions engage with one another. Truth puts the cogs in cognition, at least in its public manifestations.[4]

To use a Rylean metaphor, my view is thus that truth supplies factual dialogue with its essential esprit de corps. As the metaphor is meant to suggest, what matters is that speakers think that there is such a norm—that they take themselves to be governed by it—not that their view be somehow confirmed by science or metaphysics. Science has already done its work, in pointing out the function of the thought in the lives of creatures like us. This may suggest that a commitment to truth is like a commitment to theism, an analogy which Rorty himself draws, against Wright, in the paper with which I began. In effect, Rorty's point is that it is one thing to establish that we do employ a realist notion of truth, a normative notion stronger than justification; quite another to establish that we ought to do so. As in the case of theism, we might do better to wean ourselves of bad realist habits.

---

[4] If private cognition depends on the norms of public dialogue then truth plays the same role, at second hand, in the private sphere. This is a plausible extension of the present claim, in my view, but I will not try to defend it here

There are several important differences between the two cases, however. First, the behavioral consequences of giving up theism are significant but hardly devastating.[5] But if I am right about the behavioral role of truth, the consequences of giving up truth would be very serious indeed, reducing the conversation of mankind to a chatter of disengaged monologues.[6]

Second, it is doubtful whether giving up truth is really an option open to us. I suspect that people who think it is an option have not realized how deeply embedded the idea of truth is in linguistic practice, and therefore underestimate the extent of the required change in two ways. They fail to see how radically different from current practice a linguistic practice without truth would have to be, and they overestimate our capacity to change our practices in general to move from here to there (underestimating the practical inflexibility of admittedly contingent practices).[7]

Third, and most interestingly of all, the issue of the status of truth is enmeshed with the terms of the problem, in a way which is quite uncharacteristic of the theism case. Metaphysical conclusions tend to be cast in semantic vocabulary. Theism is said to be in error in virtue of the fact that its claims are not *true*, that its terms fail to *refer*. For this reason, it is uniquely difficult to formulate a meaningful antirealism about the semantic terms themselves. In my view the right response to this is not to think (with Paul Boghossian[8]) that we thereby have a transcendental argument for semantic realism. Without an intelligible denial, realism is no more intelligible than antirealism. The right response—as Rorty himself in any case urges—is to be suspicious of the realist-antirealist debate itself.[9] Rorty, however, ties rejection of the realist-antirealist debate to rejection of a notion of truth distinct from justification, and of the idea of representation. I think this is the wrong path to the right conclusion. We should reject the metaphysical stance not by rejecting truth and representation, but by recognizing that in virtue of the most plausible story about the function and origins of these notions, they simply do not sustain that sort of metaphysical weight.

---

[5] At least compared to the alternative.

[6] "Global *Waiting for Godot*," as a member of an audience in Dundee suggested I put it. Even more seriously, as noted above, giving up truth might silence our own "internal" rational dialogues.

[7] Jonathan Rée makes a point of this kind against Rorty: "[C]ontingencies can last a very long time. Our preoccupations with love and death may not be absolute necessities, but they are not a passing fad either, and it is a safe bet that they will last as long as we do"—"Strenuous Unbelief," *London Review of Books*, xx, 20 (October 15, 1998): 7–11, here p. 11

Concerning his own view of truth, Rorty describes himself as oscillating between Jamesian pragmatism, on the one hand, and deflationism, on the other: "swing[ing] back and forth between trying to reduce truth to justification and propounding some form of minimalism about truth."[10] My own view is neither of these alternatives, but has something in common with each. On the one hand, it is certainly some sort of minimalism about truth, but not the familiar sort that Rorty has in mind—not "Tarski's breezy disquotationalism," as he calls it (*ibid.*, p. 21). I agree with familiar disquotationalist minimalists such as W. V. Quine[11] and Paul Horwich[12] that truth is not a substantial property, about the nature of which there is an interesting philosophical issue. Like them, I think that the right approach to truth is to investigate its function in human discourse—to ask what difference it makes to us to have such a concept. Unlike such minimalists, however, I do not think the right answer to this question is that truth is merely a grammatical device for disquotation. I think that it has a far more important function, which requires that it be the expression of a norm. But like other minimalists, again, I think that there is no further question of interest to philosophy, once the question about function has been answered.

On the other hand, my view of truth is also pragmatist, for it explicates truth in terms of its role in practice. (This is also true of standard disquotational views, of course, although they ascribe the truth predicate a different role in practice.) In another sense, it conflicts with pragmatism, for it opposes the proposal that we identify truth with justification. This contrast reflects a deep tension within pragmatism. From Peirce and James on, pragmatists have often been unable to resist the urge to join their opponents in asking "What is truth?" (Indeed, the pragmatist position as a whole is often characterized in terms of its answer to this question.) Pragmatism thus turns its back on alternative paths to philosophical illumination about truth, even though these alternative paths—explanatory and genealogical approaches—are at least compatible with, if not mandated by, the pragmatist doctrine that we understand problematic notions in terms of their practical significance.

---

[8] "The Status of Content," *Philosophical Review*, XCIX, 2 (April 1990) 157–84.

[9] A realist could object that a commitment to the third norm might be useful and yet in error, but Rorty cannot. It is fair for him to object against Wright that this commitment might be like theism, because Wright takes metaphysics seriously By Wright's professed standards, then, the theism objection poses a real threat.

[10] "Is Truth a Goal of Inquiry? Donald Davidson versus Crispin Wright," see p 21.

[11] *Philosophy of Logic* (Englewood Cliffs, NJ: Prentice-Hall, 1970)

[12] *Truth* (Cambridge Blackwell, 1990).

Rorty himself is well aware of this tension within pragmatism. In "Pragmatism, Davidson, and Truth," for example (*op. cit.*), he notes that James is less prone than Peirce to try to answer the "ontological" or reductive question about truth, and suggests that Davidson may be thought of as a pragmatist in the preferable nonreductive sense.[13] As he swings between pragmatism and deflationism, then, Rorty himself is at worst only intermittently subject to this craving for an *analysis* of truth. All the same, it seems to me that he is never properly aware of the range of possibilities for nonreductive pragmatism about truth. In particular, he is not properly aware of the possibility that such a pragmatism might find itself explaining the fact that the notion of truth in ordinary use is (and perhaps ought to be, in whatever sense we might make of this) one that conflicts with the identification of truth with justification: a normative goal of inquiry, stronger than any norm of justification, of the very kind that realists about truth—opponents both of pragmatism and of minimalism—mistakenly sought to analyze. In other words, Rorty seems to miss the possibility that the right thing for the explanatory pragmatist to say might be that truth is a goal of inquiry distinct from norms of justification, and that the realist's mistake is to try to *analyze* this normative notion, rather than simply to investigate its function and genealogy. It is this latter possibility that I want to defend.

I. FALSITY AND LESSER EVILS

As I have said, I want to argue that truth plays a crucial role as a norm of assertoric discourse. It is not the only such norm, however, and a good way to highlight the distinctive role of truth is to distinguish certain weaker norms, and to imagine a linguistic practice which had those norms but not truth.[14] By seeing what such a practice lacks, we see what truth adds.

---

[13] Robert Brandom makes a similar point in "Pragmatism, Phenomenalism, and Truth Talk," *Midwest Studies in Philosophy*, XXII (1988): 75–93

There are at least two weaker norms of assertion, in addition to any distinctive norm of truth.[15] The weakest relevant norm seems to be that embodied in the principle that it is prima facie appropriate to assert that *p* only when one believes that *p*—prima facie, because of course many other factors may come into play in determining the appropriateness of a particular assertion in a particular context. Let us call this the norm of *subjective assertibility*.[16] The norm is perhaps best characterized in negative form—that is, in terms of the conditions under which a speaker may be censured for failing to meet it:

> (Subjective assertibility) A speaker is incorrect to assert that *p* if she does not believe that *p*; to assert that *p* in these circumstances provides prima facie grounds for censure, or disapprobation.

The easiest way to see that this norm has very little to do with truth is to note that it is analogous to norms which operate with respect to utterances which we do not take to be truth apt. Prima facie, it is inappropriate to request a cup of coffee when one does not want a cup of coffee, but this does not show that requests or expressions of desires are subject to a norm of truth. In effect, this norm is simply that of sincerity, and some such norm seems to govern much conventional behavior. Conventions often depend on the fact that communities censure those who break them in this specific sense, by acting in bad faith.

The second norm is that of (personal) *warranted assertibility*. Roughly, '*p*' is warrantedly assertible by a speaker who not only believes that *p*, but is *justified* in doing so. The qualification 'personal' recognizes the fact that there are different kinds and degrees of warrant or justification, some of them more subjective than others. For example, is justification to be assessed with reference to a speaker's actual evidence as she (presently?) sees it, or by some less subjective lights? For the moment, for a degree of definiteness, let us think of it in terms of subjective coherence—a belief is justified if supported by a speaker's other current beliefs. This is what I shall mean by *personal* warranted assertibility.

---

[14] For present purposes I can remain open-minded on the question as to whether such a practice is really possible. Perhaps a truth-like norm is essential to any practice which deserves to be called linguistic. At any rate, my use of the following linguistic thought experiment does not depend on denying this possibility.

[15] In what sense "weaker"? In the sense, at least, that they apply to a wider range of linguistic behaviors 'Less specialized' might be a better term.

[16] This corresponds to a common use of the term 'assertibility condition', as for example when it is said that the subjective assertibility condition for the indicative conditional 'If *p* then *q*' is a high conditional credence in *q* given *p*.

Again, this second norm is usefully characterized in negative or censure form:

> (Personal warranted assertibility) A speaker is incorrect to assert that *p* if she does not have adequate (personal) grounds for believing that *p*; to assert that *p* in these circumstances provides prima facie grounds for censure.

A person who meets both the norms just identified may be said to have done as much as possible, *by her own current lights*, to ensure that her assertion that *p* is in order. Obviously, realists will say that her assertion may nevertheless be incorrect. Subjective assertibility and (personal) warranted assertibility do not guarantee truth. To an extent, moreover, most pragmatists are likely to agree. Few people who advocate reducing truth to (or replacing truth by) a notion of warranted assertibility have personal warranted assertibility in mind. Rather, they imagine some more objective, community-based variant, according to which a belief is justified if it coheres appropriately with the other beliefs of one's community. If we call this *communal* warranted assertibility, then the point is that we can make sense of a gap between the personal and communal notions. A belief may be justified in one sense but not the other.

Pragmatists and realists may thus agree that there is a normative dimension distinct from subjective assertibility and personal warranted assertibility—an assertion may be *wrong*, despite meeting these norms. This does not yet establish that the normative standard in question need be marked in ordinary discourse. In principle, it might be a privileged or theoretical notion, useful in expert second-order reflection on linguistic practice but unnecessary in folk talk about other matters. In practice, however, there seems a very good reason why it should not remain restricted in this way. Unless individual speakers recognize such a norm, the idea that they might *improve* their views by consultation with the wider community is simply incoherent to them. (It would be as if we gave a student full marks in an exam, and then told him that he would have done better if his answers had agreed with those of other students.)

It may seem that as yet, this argument does not favor realism over pragmatism. If the normative standard an individual speaker needs to acknowledge is that of the community as a whole, there is as yet no pressure to a notion of truth beyond community-wide warranted assertibility. But what constitutes the relevant community? At any given stage, is not the relation of a given community to its possible present and future extensions just like that of the individual to her community? If so, then the same argument applies at this level. At each stage, the actual community needs to recognize that it may be wrong by the standards of some broader community.[17]

The pragmatist might now seem obliged to follow Peirce, in identifying truth with warranted assertibility in the ideal limit of inquiry. The useful thing about this limit, in this context, is that it transcends any actual community. But in my view, as I will explain below (and as Rorty in some moods already case agrees), a better move for a pragmatist is to resist the pressure to *identify* truth with anything—in other words, simply to reject the assumption that an adequate philosophical account of truth needs to answer the question "What is truth?" Better questions for a pragmatist to ask are the explanatory ones: Why do we have such a notion? What job does it do in language? What features does it need to have to play this role? And how would things be different if we did not have it?

For the moment, we have the beginnings of an answer to the last question. If we did not have a normative notion in addition to the norms of subjective assertibility and personal warranted assertibility, the idea that we might improve our commitments by seeking to align them with those of our community would be simply incoherent. I will call this the *passive* account of the role of the third norm—passive, because it does not yet provide an active or causal role for a commitment to truth. Later, I will argue that the third norm not only creates the conceptual space for argument, in this passive sense, but actively encourages speakers to participate.[18]

II. THE THIRD NORM IN FOCUS

The best way to bring the third norm into focus is again to consider its negative or censure form:

> (Truth) If not-*p*, then it is *incorrect* to assert that *p*; if not-*p*, there are prima facie grounds for censure of an assertion that *p*.

---

[17] Cf. Rorty's remark that, "[f]or any audience, one can imagine a better-informed audience"—"Is Truth a Goal of Inquiry? Donald Davidson versus Crispin Wright," see p. 22.

[18] This account has prescriptive and nonprescriptive readings. The former uses the notion of improvement full-voice, saying that if speakers are to improve their commitments, they need the idea of the third norm. But as N J. J. Smith pointed out to me, it could well be objected that the relevant notion of improvement simply presupposes the third norm, and therefore cannot provide any independent rationale for adopting it. No such circularity undermines the nonprescriptive reading, however, whose point is that because our existing conversational practice does take for granted such a notion of improvement, it thereby reveals its commitment to a third norm.

The important point is that this provides a norm of assertion which we take it that a speaker may fail to meet, even if she does meet the norms of subjective assertibility and (personal) warranted assertibility. We are prepared to make the judgment that a speaker is *incorrect*, or *mistaken*, in this sense, simply on the basis that we are prepared to make a contrary assertion; in advance, in other words, of any judgment that she fails to meet one or other of the two weaker norms.[19]

One of the reasons why this third norm is hard to distinguish from the two weaker norms of assertibility is that when we apply it in judging a fellow speaker right or wrong, the basis for our judgment lies in our own beliefs and evidence. It is not as though we are in a position to make the judgment from the stance of reality itself, as it were. I think this can make it seem as if application of this norm involves nothing more than reassertion of the original claim (in the case in which we judge it correct), or assertion of the negation of the original claim (in the case in which we judge it incorrect). Construed in these terms, our response contains nothing problematic for orthodox disquotational versions of the deflationary view, of course. Reassertion of this sort is precisely one of the linguistic activities which disquotational truth facilitates. Construed in these terms, then, there is no need for truth to be a distinct *norm*.

Our response is not merely reassertion, or assertion of the negation of the original claim, however. If it were, it would involve no commendation or criticism of the original utterance. This nonnormative alternative is hard to see, I think, because the norm in question is so familiar and so basic. As a result, it is difficult to see the immense difference the norm makes to the character of disagreements. But it comes into focus, I think, if we allow ourselves to imagine a linguistic practice which allowed reassertion and contrary assertion, but without this third normative dimension. What we need to imagine, in other words, is a linguistic community who use sentences to express their beliefs, and have a purely disquotational truth predicate, but for whom disagreements have no normative significance, except insofar as it is related to the weaker norms of assertibility.

This imaginative project is not straightforward, of course. Indeed, it is not clear that it is entirely coherent. If there is a third norm of the kind in question, is not it likely to be constitutive of the very notions of assertion and belief? If so, what sense is there in trying to imagine an assertoric practice which lacked this norm?

---

[19] Note the contrast with Rorty's cautionary use of "true." In that use, we say of a claim that we take to be well justified that it might not be true. In the present use, we say of a claim that we might even allow to be well justified by its speaker's own lights that it is not true. It is the difference between mere caution and actual censure

Well, let us see. What we need is the idea of a community who take an assertion—or rather the closest thing they have to what we call an assertion—to be *merely* an expression of the speaker's opinion. The relevant idea is familiar in the case of expressions of desires and preferences. It is easy to imagine a community—we are at least close to it ourselves—who have a language in which they give voice to psychological states of these kinds, not by *reporting* that they hold them (which would depend on assertion), but directly, in conventional linguistic forms tailored specifically for this purpose.

Think of a community who use language primarily for expressing preferences in restaurants, for example. (Perhaps the development of such a restricted language from scratch is incoherent, but surely we might approach it from the other direction. Imagine a community of dedicated lunchers, whose language atrophies to the bare essentials.) In this community we would expect a norm analogous to subjective assertibility: essentially, a normative requirement that speakers use these conventional expressions sincerely. Less obviously, such a practice might also involve a norm analogous to personal warranted assertibility. In other words, expressed preferences might be censured on the grounds that they were not well founded, by the speaker's own lights (for example, on the grounds they did not cohere with the speaker's other preferences and desires). In this practice there need be no place for a norm analogous to truth, however—no idea of an objective standard, over and above personal warranted assertibility, which preferences properly aim to meet.

At least to a first approximation, we can imagine a community who treat expressions of beliefs in the same way. They express their beliefs—that is, let us say, the kind of behavioral dispositions which we would characterize as beliefs—by means of a speech act we might call the *merely-opinionated assertion* (MOA, for short). These speakers—"Mo'ans," as I called them in another article[20]—criticize each other for insincerity and for lack of coherence, or personal warranted assertibility. But they go no further than this. In particular, they do not treat a disagreement between two speakers as an indication that, necessarily, one speaker or the other is mistaken—in violation of some norm. On the contrary, they allow that in such a case it may turn out that both speakers have spoken correctly, by the only two standards the community takes to be operable. Both may be sincere, and both, in their own terms, may have good grounds for their assertion.[21]

---

[20] "Three Norms of Assertibility, or How the MOA Became Extinct," *Philosophical Perspectives*, XII (1988): 241–54. The present section and the next draw significantly on that paper.

A speech community of this imagined kind could make use of a disquotational truth predicate, as a device to facilitate agreement with an expression of opinion made by another speaker. 'That's true' would function much like 'Same again', or 'Ditto', used in a bar or restaurant. Just as 'Same again' serves to indicate that one has the same preference as a previous speaker, 'That's true' would serve to indicate that one holds the same opinion as the previous speaker. The crucial point is that if the only norms in play are subjective assertibility and personal warranted assertibility, introducing disquotational truth leaves everything as it is. It does not import a third norm.

The difficulty we have in holding on to the idea of such a community stems from our almost irresistible urge to see the situation in terms of our own normative standards. There really is a third norm, we are inclined to think, even if these simple creatures do not know it. If two of them make incompatible assertions then one of them must be objectively incorrect, even if by their lights they both meet the only norms they themselves recognize. (I think even pragmatists will be inclined to say this, even though they want to equate the relevant kind of incorrectness not with falsity but with lack of some kind of justification more objective than that of personal warrant.) But the point of the story is precisely to bring this third norm into sharp relief, and hence I am quite happy to allow challenges to the story on these grounds, which rely on the very conclusion I want to draw. *For us*, there is a third norm. But why is that so? Where does the third norm come from? What job does it do—what difference does it make to our lives? And what features must it have in order to do this job?

III. WHAT DIFFERENCE DOES THE THIRD NORM MAKE?

Let us return to the Mo'ans, and their merely-opinionated assertions. Recall that Mo'ans use linguistic utterances to express their "beliefs," as well as other psychological states, such as preferences and desires. Where they differ from us is in the fact that they do not take a disagreement between two speakers in this belief-expressing linguistic dimension to indicate that one or another speaker must be at fault. They recognize the possibility of fault consisting in failure to observe one of the two norms of subjective assertibility or personal warranted assertibility, but lack the idea of the third norm, that of truth itself. This shows up in the fact that by default, disagreements are of a no-fault kind, in the way that expressions of different preferences often are for us.

---

[21] As I noted earlier, my use of this example does not depend on the claim that such a linguistic practice be possible. It is doubtful whether notions such as belief, assertion, and opinion are really load bearing, in the imagined context. Much of the effect of the example could be achieved in another way, however, by imposing suitable restrictions on real linguistic practices—by imagining self-imposed restrictions on what we are allowed to say. One way to approach the Mo'an predicament from our own current practice would be to adopt the convention that whenever we would ordinarily assert '*p*', we express ourselves instead by saying 'My own opinion is that *p*'.

What does it take to add the third norm to such a practice? Do the Mo'ans need to come to believe that there is a substantial property that the attitudes they use MOAs to express may have or lack—perhaps the property of corresponding to how things are in the world, perhaps that of being what their opinions are fated to converge on in the long run? Does adoption of the third norm depend on a piece of folk metaphysics of this kind? Not at all, in my view. The practice the Mo'ans need to adopt is simply that whenever they are prepared to assert (in the old MOA sense) that *p*, they also be prepared to ascribe fault to anyone who asserts not-*p*, independently of any grounds for thinking that that person fails one of the first two norms of assertibility. Perhaps they also need to be prepared to commend anyone who asserts that *p*, or perhaps failure-to-find-fault is motivation enough in this case. At any rate, what matters is that disagreement itself be treated as grounds for disapproval, as grounds for thinking that one's interlocutor has fallen short of some normative standard.

At this point it is worth noting what may seem a serious difficulty. If the Mo'ans do not *already* care about disagreements, why should they care about disagreements about normative matters? Suppose that we two are Mo'ans, that you assert that *p*, and that I assert that not-*p*. If this initial disagreement does not bother me, why should it bother me when—trying to implement the third norm—you go on to assert that I am "at fault," or "incorrect"? Again, I simply disagree; and if the former disagreement does not bite then nor will the latter. And if what was needed to motivate me to resolve our disagreement was *my* acceptance that I am "at fault," then motivation would always come too late. If I accept this at all, it is only after the fact—after the disagreement has been resolved in your favor.

To get the sequence right, then, I must be motivated by your disapproval itself. This is an important point. It shows that if there could be an assertoric practice which lacked the third norm, we could not add that norm simply by adding a normative predicate. Insofar—so very far, in my view—as terms such as 'true' and 'false' carry this normative force in natural languages, they must be giving voice to something more basic: a fundamental practice of expressions of attitudes of approval and disapproval, in response to perceptions of agreement and disagreement between expressed commitments. I will return to this point, for it is the basis of an important objection to certain other accounts of truth.

Imagine for the moment that the Mo'ans could add the third norm by adding a normative predicate, or pair of predicates ('correct' and 'incorrect', say). What would be the usage rule for these predicates? Simply that one be prepared to assert that *p* is correct if and only if one is prepared to assert that *p*, and to assert that *p* is incorrect if and only if one is prepared to assert that not-*p*. In other words, the usage rule is something very close to the disquotational schema ('*p*' is true if and only if *p*). As a result, the present proposal, that the truth predicate is an explicit expression of the third norm, already seems well on the way to an explanation of the disquotational functions of truth. We have already noted that the converse argument does not go through. A practice which lacked the third norm could still make use of a disquotational truth predicate.[22]

For the moment, we are interested in the function of the third norm. Why might the invention of such a norm be useful? What distinctive job does it do? We already have one answer to the latter question, and hence possibly to the former, in the passive account. Without a norm stronger than that of warranted assertibility *for me*, or *for us*, the idea of improving *my*, or *our*, current commitments would be incoherent. The third norm functions to create the conceptual space for the idea of further improvement. To do this job, we need a norm stronger than that of warranted assertibility for any *actual* community. (Of course, this does not yet show that we need something more than Peircean ideal assertibility, but one thing at a time.)

We can do better than the passive account, however. The third norm does not just hold open the conceptual space for the idea of improvement. It positively encourages such improvement, by motivating speakers who disagree to try to resolve their disagreement. Without the third norm, differences of opinion would simply slide past one another. Differences of opinion would seem as inconsequential as differences of preference. With the third norm, however, disagreement automatically becomes normatively loaded. The third norm makes what would otherwise be no-fault disagreements into unstable social situations, whose instability is only resolved by argument and consequent agreement—and it provides an immediate incentive for argument, in that it holds out to the successful arguer the reward consisting in her community's positive evaluation of her dialectical position. If reasoned argument is generally beneficial—beneficial in some long-run sense—then a community of Mo'ans who adopt this practice will tend to prosper, compared to a community who do not.

---

[22] A defender of the disquotational view might argue that although there is a third norm, it is not the function of the truth predicate to express it. This will be a difficult position to defend, however. If any predicate—'correct', for example—expresses the third norm, then that predicate will function as a disquotational predicate, for the reason just mentioned. Hence it will have been pointless to maintain that true itself is not normative. So the disquotationalist needs to claim that the third norm is not expressed at all in this predicative form, and that seems implausible

I will call this the *active* account of the role of the third norm. In effect, it contends that the fact that speakers take their belief-expressing utterances to be subject to the third norm plays a causal, carrot-and-stick role in encouraging them to settle their differences, in cases in which initially they disagree. The force of these carrots and sticks should not be overstated, however. In any given case, we are free not to give voice to our third-norm-grounded disapproval. If we do express it, the speakers with whom we disagree are free not to rise to the bait. Many factors may determine what happens in any particular case. My claim is simply that the third norm adds something new to the preferential mix. In particular, it gives rise to a new preferential pressure toward resolution of the disagreement in question—a pressure which would not exist in its absence, which does not exist for the Mo'ans, and which could not exist for us, if we did not care in general about the approval and disapproval of our fellows. The third norm depends on the fact that (to varying extents in varying circumstances) we do care about these things. It exploits this fact about us to make disagreements matter, in a way in which they would not otherwise matter. But the third norm does not come for free, with a general disposition to seek the approval of our fellows. What we have but the Mo'ans lack is an additional, special purpose, disposition: the disposition to disapprove of speakers with whom we disagree. This disposition is the mark of the third norm.

As in the case of the passive account of the role of the third norm, we need to be careful that this active account does not viciously presuppose the very notions for which it seeks to account. The notion of disagreement requires particular care. For one thing, recognition that one differs from a previous speaker must take some form more basic than the belief that he or she has said something "false," for otherwise there could not be a convention of applying this normative predicate when one perceives that one differs. For another, there is an important sense in which, on the proposed account, it is the practice of applying the third norm which creates the disagreement, where initially there was mere difference. Properly developed, the view seems likely to be something like this. There is a primitive incompatibility between certain behavioral commitments[23] of a single individual, which turns on the impossibility of both doing and not doing any given action *A*—both having and not having a cup of coffee, for example. All else—both the public perceived incompatibility of "conflicting" assertions by different speakers, and the private perceived incompatibility essential to reasoning—is by convention, and depends on the third norm.

Obviously, much more needs to be said about this. At another level, much also needs to be said about the possible advantages of such a mechanism for resolving differences—about its long-run advantages, for example, both compared to the case in which there is no such mechanism and compared to the case in which there is some different mechanism, such as deference to social rank. For immediate purposes, however, my claim does not depend on this latter work. For the present, my claim is simply that truth does play the role of this third norm, in providing the friction characteristic of factual dialogue as we know it. (I also claim, roughly, that this is perhaps the most interesting fact about truth, from a philosophical perspective.) In principle, this claim could be true, even though the practice in question was not advantageous. In principle, truth, and with it dialogue, could turn out to be a bad thing for the species, biologically considered.[24] No matter; it would still be true that we would not have understood truth until we understood its role in this debilitating practice.

---

[23] This is another place where circularity threatens. We need to be sure that the psychological states mentioned at this point are not thought of as already "factual" or "representational" in character, in a way which presupposes truth. In so far as it is truth-involving, the "factual" character of the domain in question needs to be part of the explanandum—something that emerges from, rather than being presupposed by, the pragmatic account of the origins and consequences of "truth talk." In my view, one of the attractive features of this approach is that it offers the prospect that the uniformity of "factual," truth-involving talk might be compatible with considerable plurality in the nature and functions of the underlying psychological states. It thus offers an attractive new form for expressivist intuitions. Cf. my *Facts and the Function of Truth* (New York. Blackwell, 1988), chapter 8; "Metaphysical Pluralism," this JOURNAL, LXXXIX, 8 (August 1992) 387–409, §IV, and "Immodesty without Mirrors: Making Sense of Wittgenstein's Linguistic Pluralism," in Max Kölbel and Bernhard Weiss, eds., *Wittgenstein's Lasting Significance* (New York: Routledge, forthcoming).

[24] Even if not dangerous on its own, the third norm might become so in combination with some particularly deadly source of intractable disagreements, such as religious commitment More generally, the thought that argument is sometimes dangerous suggests a link between the concerns of this paper and the motivations of the Pyrrhonian skeptics. On the present view of truth, the question whether we could get by without truth seems closely related to that as to whether we could live as thoroughgoing Pyrrhonian skeptics

Is talk of dialogue really essential here? Could not we say simply that the third norm is what distinguishes a genuinely assertoric linguistic practice from the "merely opinionated" assertoric practice of the Mo'ans? The distinguishing mark of genuine assertion is thus that by default, difference is taken as a sign of *fault*, of breach of a normative standard.

It would not strictly be incorrect to say this, in my view, but it ought to seem unsatisfying, by pragmatist lights. A pragmatist is interested in the practical significance of the notions of truth and falsity, in the issue of what difference possession of these notions makes to our lives. According to the view just suggested, the answer will be something like this. The difference that truth and falsity make is that they make our linguistic practice genuinely assertoric, rather than Mo'an. "I see that," the pragmatist will then say, "But what practical difference does *that* difference make, over and above the obvious difference—that is, over and above the fact that we approve and disapprove of some of our fellow speakers on occasions on which we would not otherwise do so?"

My own answer to the new question is that these habits of approval and disapproval tend to encourage dialogue, by providing speakers with an incentive to resolve disagreements. It is true that at this point the pragmatist's question—"What difference does *that* make?"—can be (indeed, should be) asked all over again. The importance I have here attached to dialogue rests in part on the gamble that this question will turn out to have an interesting answer, in terms of the long-run advantages of pooled cognitive resources, agreement on shared projects, and so on. But not entirely. Dialogue seems such a central part of our linguistic and social lives that the difference between a world without dialogue and our world is much greater than *merely* the difference between MOAs and genuine assertions. So even if it were to turn out that the development of dialogue had been an historical accident, of no great value to the species biologically considered, it would still be true that the most interesting behavioral consequence of the third norm would be dialogue, and not merely the more-than-merely-Mo'an assertion which makes dialogue possible.[25]

Recall that I began by challenging Rorty's claim that no behavioral consequences flow from a distinction between justification and truth.

---

[25] This point would acquire new and even stronger force, if it were to be established that private cognition rests on the norms of public dialogue, in the way suggested in footnote 4.

In one sense, my challenge does indeed amount to pointing out that the third norm—a notion of truth stronger than justification—brings with it the following behavioral difference: a disposition to criticize, or at least disapprove of, those with whom one disagrees. But if this were all the challenge amounted to, Rorty would be entitled to reply that of course there is this difference, but that this difference makes no interesting further difference. Hence the importance of dialogue, in my view, which turns a small difference in normative practice into a big difference in the way in which speakers engage with one another (and thereby ensures that Rorty's claim fails in an interesting rather than an insignificant way).

IV. PEIRCE REGAINED?

Now to the question deferred above. Does the third norm need to be other than a more-than-merely-personal notion of justification? In particular, could not it be a Peircean flavor of ideal warranted assertibility? I have several responses to this suggestion.

First, I think that the proposal is mistakenly motivated. As I said in the introduction, I think it stems from the tendency, still too strong in Peirce, to ask the wrong question about truth. If we think that the philosophical issue is "What is truth?" then naturally we will want to find an answer—something with which we may identify truth. Then, given standard objections to metaphysical answers, it is understandable that Peirce's alternative should seem attractive. But the attraction is that of methadone compared to heroin. Far better, surely, from a pragmatist's point of view, to rid ourselves of the craving for analysis altogether. To do this, we need to see that the basic philosophical needs that analysis seemed to serve can be met in another mode altogether: by explanation of the practices, rather than reduction of their objects. (Moreover, the explanatory project has the potential to allow us realist truth without the metaphysical disadvantages. The apparent disadvantages of realist truth emerge in the light of the reductive project, for it is from this perspective that it seems mysterious what truth could be. If we no longer feel obliged to ask the question, we will not be troubled by the fact that it is so hard to answer. We lose the motivation for seeking something else—something less "mysterious" than correspondence—with which to identify truth.)

"I accept all that," the pragmatist might say. "Nevertheless, perhaps it is true of the notion of truth (as we find it in practice), that it is identical (in some interesting sense) to ideal warranted assertibility. Should you not therefore allow, at least, the possibility that a Peircean account is the correct one?"

Two points in response to this: the first, an old objection, is that it is very unclear what the notion of the ideal limit might amount to, or even that it is coherent. For example, could not actual practice be improved or idealized in several dimensions, not necessarily consistent with one another? In this sense, then, the Peircean pragmatist seems a long way from offering us a concrete proposal.[26]

The second point—also an old point, for as Hilary Putnam observes, it is essentially the naturalistic fallacy[27]—concerns the nature of the proposed identification of truth with ideal warranted assertibility. Truth is essentially a normative notion. Its role in making disagreements matter depends on its immediate motivational character. Why should ideal warranted assertibility have this character? If someone tells me that my beliefs are not those of our infinitely refined future inquirers, why should that bother me? My manners are not those of the palace, but so what? In other words, it is hard to see how such an identification could generate the immediate normativity of truth.[28] (It seems more plausible that we begin with truth and define the notion of the ideal limit in terms of it: what makes the limit ideal is that it reaches truth. This does not tell us how and why we get into this particular normative circle in the first place.)

I have not yet mentioned what seems to me to be the most telling argument against the pragmatist identification of truth with warranted assertibility (in Peircean form or otherwise). It often seems to be suggested (by Rorty himself, among others—see the quotes with which I began), that instead of arguing about truth, we could argue about warranted assertibility. This seems to me to miss a crucial point. Without truth, the wheels of argument do not engage; disagreements slide past one another. This is true of disagreements about any matter whatsoever. In particular, it is true of disagreements about warranted assertibility. If we did not already have truth, in other words, we simply could not argue about warranted assertibility. For we could be aware that we have different opinions about what is warrantedly assertible, without that difference of opinion seeming to matter. What makes it matter is the fact that we subscribe to a practice according to which disagreement is an indication of culpable error, on one side or other; in another words, that we already take ourselves to be subject to the norms of truth and falsity.[29]

---

[26] As Rorty notes (in "Pragmatism, Davidson, and Truth," see p. 130), Michael Williams makes a point of this kind: "[W]e have no idea what it would be for a theory to be ideally complete and comprehensive ..or of what it would be for inquiry to have an end"—"Coherence, Justification, and Truth," *Review of Metaphysics*, xxxiv (1980–81): 243–72, see p. 269.

[27] *Meaning and the Moral Sciences* (New York: Routledge, 1978), p. 108.

[28] It may seem that this argument begs the question against the pragmatist, by assuming that there is an epistemologically relevant gap between ideal warranted assertibility and truth. But the issue is not whether we need some norm in addition to ideal warranted assertibility, but whether ideal warranted assertibility itself could be immediately normative, in the way in which truth is. No one disputes that the manners of the palace are normative for those who live there—that is what it is to be manners—but it is an open question whether they are or should be normative for the rest of us. Similarly for ideal assertibility, except that in this case no one lives at the limit, so that there is no one for whom the question is not open

The crucial point is thus that assertoric dialogue requires an intolerance of disagreement. This needs to be present already in the background, a pragmatic presupposition of judgment itself. I am not a maker of assertions, a judger, at all, unless I am already playing the game to win, in the sense defined by the third norm. Since winning is already characterized in terms of truth, the idea of a conversational game with some alternative point is incoherent. It is like the idea of a game in which the primary aim is to compete—this idea is incoherent, because the notion of competition already presupposes a different goal.[30]

There is a connection here with an old objection to relativism, which tries to corner the relativist by asking her whether she takes her own relativistic doctrine to be true, and if so in what sense. The best option for the relativist is to say that she takes the doctrine to be true in the only sense she allows, namely, the relativistic one. When her opponent replies, "Well, in that case you should not be troubled by the fact that I disagree, because you recognize that what is true for me need not be true for you, and vice versa," the relativist has a reply. She can argue that truth is relative to communities, not to individual speakers, and hence that disagreements do not necessarily dissolve in this way.

---

[29] I noted above that the same point applies to the normative predicates themselves. If we were not already disposed to take disagreement to matter, we could not do so simply by adding normative predicates, for disagreement about the application of those predicates would be as frictionless as disagreement about anything else. My claim is thus that the notions of truth and falsity give voice to more primitive implicit norms, which themselves underpin the very possibility of "giving voice" at all. In effect, the above argument rests on the observation that this genealogy cannot be reversed: if we start with a predicate—warrantedly assertible or any other—then we have started too late. (I suspect that by "giving voice," I mean something close to "making explicit," in Brandom's sense.)

[30] Here, incidentally, we see the essential flaw in the pious sentiments of Grantland Rice (1880–1954):

For when the One Great Scorer comes,
To write against your name,
He marks—not that you won or lost—
But how you played the game.

The One Great Scorer might assign marks on this basis, for divine purposes. *Pace* Rice, however, we could not play the game in question with such marks as our primary goal, for then it would be a different game altogether.

My pragmatist opponents fare less well against an analogous argument, I think. The basic objection to their position is that in engaging with me in argument about the nature of truth (as about anything else), they reveal that they take themselves to be subject to the norm whose existence they are denying. If they did not take themselves to be subject to it, they would be in the same boat as the Mo'ans, with no reason to treat the disagreement between us as a cause for concern. They affirm *p*, I affirm not-*p*, but by their lights, this should be like the case in which they say "Yes" and I say "No," in answer to the question "Would you like coffee?" (This is what it should be like even if *p* is of the form '*q* is warrantedly assertible'.) The disagreement simply would not bite.

V. TRUTH AS CONVENIENT FICTION?

The third norm thus requires a notion of truth that differs from justification, even of a Peircean ideal variety. In this sense, then, the present account is realist rather than pragmatist about truth. In another sense, however, the view surely seems antirealist. After all, I have argued that what matters is that speakers take there to be a norm of truth, not that there actually be such a norm, in some speaker-independent sense. Is this not antirealism, or more precisely, in the current jargon, a form of *fictionalism* about truth?

If so, could this be a satisfactory outcome? If truth does play the role I have claimed for it in dialogue, would not the realization that it is a fiction undermine that linguistic practice, by making it the case that we could no longer consistently feel bound by the relevant norms?

Let us call this objection the threat of *dialogical nihilism*. In my view, it is not a practical threat. I think that in practice we find it impossible to stop caring about truth. This is not an argument for realism, of course. The discovery that our biological appetites are not driven by perception of pre-existing properties—the properties of being tasty, sexually attractive, or whatever—does not lessen the force of those appetites, but no one thinks that this requires a realist view of the properties concerned. Even if nihilism were a practical threat, this would not be reason for thinking that the claim that truth is a fiction is *false*, by the lights of the game as currently played. It might be a pragmatic reason for keeping the conclusion quiet, but that is a different matter altogether (especially according to my realist opponents).

So even if the present view is correctly characterized as a form of fictionalism about truth, the nihilism objection is far from conclusive. But are the labels 'fictionalism' or 'antirealism' really appropriate? The need for caution stems from the fact that this approach to truth threatens to deprive both sides of the realism-antirealism debate of conceptual resources on which the debate seems to depend. As I noted earlier, the relevant metaphysical issues tend themselves to be framed in terms of truth, and related notions. A theory is said to be in error if its claims are not *true*, or if its terms fail to *refer*, for example. So the issue of the status of truth is here enmeshed with the terms of the problem, in a way which is quite uncharacteristic of metaphysical issues about other notions. As a result, it may be impossible to formulate a meaningful antirealism or fictionalism about the semantic terms themselves. This does not mean that we have to be realists about semantic notions, but only that if we are not realists we should be cautious about calling ourselves antirealists (or fictionalists), if these categories presuppose the very notions we want to avoid being realist about.

This may sound like an impossible trick, but in fact the kind of distinction we need is familiar elsewhere. It is the distinction between someone who "talks god talk" and espouses atheism, and someone who rejects the theological language game altogether (on Carnapian pragmatic grounds, say). These are two very different ways of rejecting theism. In the present case, the point is that we may consistently reject semantically-grounded realism about the semantic notions themselves, so long as we do so by avoiding theoretical use of semantic notions altogether, rather than by relying on those notions to characterize our departure from realism. (Why "theoretical use"? Because there is nothing to stop us continuing to use these semantic notions in a deflationary or disquotational sense.)

It might be suggested that we can sidestep this difficulty altogether by casting the relevant metaphysical issues in ontological rather than semantic terms. On this view, the relevant issue is whether truth exists, not whether (some) truth ascriptions are true. Against this suggestion, however, it is arguable that the relevant metaphysical issues arise initially from data concerning human linguistic usage, and only become metaphysical in the light of substantial semantic assumptions about the functions of the language concerned—for example, that it is truth conditional, or referential, in function. If so, then truth is once again enmeshed with the terms of the problem. And even if we concede the possibility of the ontological shift, the authority of Quine, Carnap and others may perhaps be invoked in support of a deflationary attitude to ontology, with the result that the realist-antirealist issue still dissolves.[31]

These issues are complex, and deserve a much more detailed examination than I can give them here. For present purposes, I simply flag the following as a possible outcome (of considerable plausibility, in my view). In common with other deflationary approaches to truth, the present account not only rejects the idea that there is a substantial metaphysical issue about truth (a substantial issue about the truthmakers of claims about truth, for example). Because it is about truth, it also positively prevents "reinflation." In other words, it seems to support a general deflationary attitude to issues of realism and antirealism. If so, then deflationism about truth is not only not to be equated with fictionalism, but tends to undermine the fictional-nonfictional distinction, as applied in the metaphysical realm.[32]

As I noted at the beginning, the present account of truth is hard to find on contemporary maps. In part, as should now be clear, this is because it combines elements not normally thought to be compatible. In one sense it is impeccably pragmatist, for example, for it appeals to nothing more than the role of truth in linguistic practice. Yet it rejects the pragmatist's *ur*-urge, to try to identify truth with justification. Again, it defends a kind of truth commonly seen as realist, but does so from a pragmatist starting point, without the metaphysics that typically accompanies such a realist view of truth. So in thinking about how to characterize this account of truth, we should be sensitive to the possibility that our existing categories—fictionalism, realism, and perhaps pragmatism itself—may need to be reconfigured. If so, then putting the position on the map is not like noticing a small country (Lichtenstein, perhaps) that previously we had overlooked. It is more like discovering a geographical analogue of the platypus, a region which our pre-existing cartographical conventions seemed a priori to disallow.

I began with Rorty's claim that the distinction between justification and truth makes no difference in practical life, no difference to our "decisions about what to do." Rorty regards a commitment to a notion of truth stronger than justification as a relic of a kind of religious deference to external authority. He recommends that just as we have begun to rid ourselves of theism, we should rid ourselves of the "representationalist" dogma that our beliefs are answerable to standards beyond ourselves. For Rorty, then, realist truth is a quasi-religious myth, which we would do better without.

---

[31] I defend the first of these two options in "Naturalism without Representationalism," in Mario de Caro and David Macarthur, eds., *Naturalism in Question* (Cambridge: Harvard, 2003) and the second in "Metaphysical Pluralism", and "Naturalism and the Fate of the M-worlds," *Proceedings of the Aristotelian Society*, Supp. Vol LXXI (1997). 247–67

[32] Rorty often says that he wants to walk away from realist–antirealist disputes. In other words, he does not think that there is an interesting philosophical question as to whether our commitments "mirror" reality The above argument suggests that like other deflationists about truth, I have reason to follow Rorty in walking away from these issues. (In particular, my defense of truth over justification does not force me to stay.)

Despite my reservations about the fictionalist label, I have agreed that truth is in some sense a myth, or at least a human creation.[33] But I have denied that in providing a norm stronger than justification, a commitment to truth makes no behavioral difference. On the contrary, I have argued, it plays an essential role in a linguistic practice of great importance to us, *as we currently are*. It is not clear whether we could coherently be otherwise, whether we could get by without the third norm. If so, however, then the result would be a very different language game. My main claim is that we have not understood truth until we understand its role in the game we currently play.

HUW PRICE

University of Edinburgh
University of Sydney

---

[33] In the light of the argument above, this is a point more about the genealogy than about the reality of truth

A SPECTRAL REFLECTANCE DOTH NOT A COLOR MAKE*

According to color realists, the colors we experience are typically properties of the surfaces of physical objects. It is a curious sociological fact that many philosophers, but very few visual scientists, are color realists. This is presumably because philosophers, unlike visual scientists, are much concerned to sustain what they take to be a common-sense view of colors. Perhaps some philosophers are still haunted by the ghosts of sense data, and fear that once colors are banished from the physical world, they will take up refuge in what Gilbert Ryle once called "the dust-bin of the mind," nevermore to be dislodged.

Prominent among those who attempt to keep the chromatic barbarians outside the gates are those color realists who maintain that colors are to be identified with surface spectral reflectances. Recent versions of reflectance theory have been defended by Alex Byrne and David Hilbert in their article, "Color Realism and Color Science,"[1] as well by Michael Tye in his book, *Consciousness, Color, and Content*.[2] The present article will be devoted to an examination of Tye's attempt to refute the arguments that color scientists and their philosopher camp-followers have adduced against color realism, but most of it applies as well to Byrne and Hilbert's position.

We shall evaluate Tye's defense of color realism by measuring it against two principles to which color realists ought to subscribe. The first principle is that an object's having a given color is a matter of fact. To adhere to this principle is to be a realist about color. According to the second principle, it is normally possible to determine what color a thing has by looking at it. Let us call this the *just by looking* principle. To reject it would be to deny that we have epistemic access to colors, unless it would be possible to find out what colors things have by some other means. But by what other means? We can of course measure surface spectral reflectances. Whether colors are surface spectral reflectances is the point at issue, however. Even if colors are surface spectral reflectances, there would still be the question of

---

*I wish to thank Rolf Kuehni, Brian McLaughlin, and Sydney Shoemaker for their helpful comments on earlier drafts of this article.

[1] *Behavioral and Brain Sciences* (forthcoming).

[2] Cambridge: MIT, 2000.

0022-362X/03/0004/191–202

which reflectances are which colors. And how could that be determined without appealing to what colors they make things appear to have?

Let us begin by considering the phenomenon of approximate color constancy. Tye says,

> The fact that objects appear to retain the same color through a wide variety of changes in illumination conditions (though certainly not all) strongly suggests that colors are illumination-independent properties of those objects. The simplest, most straightforward explanation of color constancy is that the surfaces of colored objects have features that remain the same as the illumination conditions change—features that are represented in our color experiences and that are responsible for the sameness in their phenomenal character (*ibid.*, p 147).

Indeed, many objects that we typically see as colored do have such illumination-independent properties, namely, their surface spectral reflectances, and an important part—but not the only part—of the function of our color-vision system is to track them. But before rashly concluding that spectral reflectances are the very colors that we see, we ought to look in more detail at the mechanisms that the eye uses to do the tracking, as well as the accuracy with which it accomplishes this task.[3]

Because the eye contains only three types of photopigment,[4] it has but three degrees of freedom with which to represent light spectra. To disentangle the spectrum of the illuminant from the spectrum of the surface under a wide variety of conditions, it would have to have five. Perfect color constancy is therefore impossible, and the eye must rely on a number of tricks to discount the illuminant as well as it does.

---

[3] Most of the empirical material on which the subsequent discussion draws may be found in the following sources.

(1) P K. Kaiser and R. M. Boynton, *Human Color Vision*, Second Edition (Washington, DC: Optical Society of America, 1996); see especially chapter 11.
(2) D. Jameson and L. M. Hurvich, "Essay Concerning Color Constancy," reprinted in Byrne and Hilbert, eds., *Readings on Color*, Volume 2 (Cambridge: MIT, 1997), pp 171–98
(3) C. L. Hardin, *Color for Philosophers: Unweaving the Rainbow*, Expanded Edition (Indianapolis: Hackett, 1993).

There are some excellent illustrations of simultaneous contrast and the conditions for the appearance of brown on the web page of Dale Purves's lab. www.purveslab net

[4] It is now known that many people, predominantly females, have more than three types of photopigments in their retinas. Because these seem not to involve additional opponent channels, however, the conclusion of this argument is at least approximately correct.

Two of the tricks are of particular interest to us here. The first is adaptation, in which the eye recalibrates its neutral point. In the case of achromatic adaptation, it approximately adjusts its middle gray to the average intensity of the light. Since the eye's inherent dynamic range is orders of magnitude less than the dynamic range of the illuminants with which it must deal, it foregoes responding to the absolute intensity of the illuminant in favor of responding to intensity variations within the scene. Analogously, in the chromatic domain, the eye reacts to a strong chromatic light—one that looks red, for example—by losing sensitivity to the relevant portion of the spectrum, increasing its relative response to red's opponent color, green. The tinge of color that one sees across a scene when one removes a pair of sunglasses reflects chromatic adaptation to the filtered spectrum of the lenses. As the eye readapts, the tinge disappears, to be replaced by another, perhaps less chromatic one. As we move through it, the light patterns of a natural scene are constantly changing, so the eye is constantly adapting. Since adaptation is neither instantaneous nor complete, the colors that we see almost always have an adaptive component. For the relatively slow transition from one natural phase of daylight to another, adaptation is gradual and the adaptive shortcomings are often small. However, in relatively rapid transitions from natural light to artificial light, or from one sort of artificial light to another, the effect of adaptation is significant, although the change in color experience that adaptation induces frequently passes unnoticed. For example, a paper that looks magenta in north daylight looks yellowish red when it is seen after a minute in incandescent light. After ten minutes in incandescent light, this paper will seem to be almost a unique red

The second trick that the eye—more strictly, the eye-brain—employs to approximate color constancy is to compute the color of a region of a scene by comparing it with other regions in the same scene. The algorithm the eye-brain employs to do this is still unknown, although several modestly successful ones have been proposed. Such algorithms typically compute the ratios that each region bears to its neighbor along the respective dimensions of lightness, greenness/redness, and blueness/yellowness. When the ratios are degenerate, as they are in the case of a single patch of color against a completely dark background, color constancy is zero. That means that the eye-brain totally confounds the spectrum of the illuminant with the reflectance spectrum of the patch, which is then seen as self-luminous. Comparison of visual regions is thus absolutely essential to discounting the illuminant and achieving relative stability of object color appearance. In addition to ratio computations, the eye-brain uses other strategies to assign a color to a surface, including the surface's orientation with respect to other surfaces. All of these strategies employ information about visual regions outside the surface in question.

Now we are prepared to examine Tye's defense of color realism against the objections of some color scientists. The first objection is that, "We sometimes see physically identical objects or spectral arrays as having different colors," and the second, the phenomenon of metamerism, is that, "We routinely see physically different spectral arrays as having the same color." Considering the second objection first, Tye concludes that it only shows that, "The color a surface has is not one and the same as the wavelength of the light it reflects under any particular illumination" (*op. cit.*, p. 150). Indeed, if the color of a surface is to be identified with its spectral reflectance, this certainly follows. But how would this accord with the principle of just by looking? Tye seems to accept something like this principle, for he invokes, "The reasonable assumption that, *ceteris paribus*, our color experiences under normal viewing conditions are veridical" (*op. cit.*, p. 160). If this is so, there must be some illuminant or range of illuminants—the "normal" ones—under which objects seem, *ceteris paribus*, to have the color that they actually do have. But given any particular illuminant and any particular observer, one can always find a pair of physical samples that will match for that observer under that illuminant, yet fail to match for that observer under other illuminants. In fact, the Macbeth Company once manufactured a metameric slide rule that permits one to find such a pair of samples for each of a wide range of illuminants. So should not color realists not only claim that there are viewing conditions that are veridical, but be prepared to specify which ones are veridical, or at least tell us how we could in principle recognize such conditions when we encounter them? A color realist might decline to shoulder this burden, appealing to the example of a mathematician who proves the existence of a set without having to tell us how we might find it. This might be acceptable if color realists actually had a proof that such veridical viewing conditions exist. That they exist is frequently asserted, but it should go without saying that assertions do not count as proofs.

To meet the first objection, that we sometimes see physically identical objects or spectral arrays as having different colors, Tye must deal with the phenomenon of simultaneous color contrast, in which the perceived color of a surface is affected by the background against which it is seen. He considers two options. The first is to hold that, "The color of an object is not always just a function of its intrinsic properties.... The color a surface has is relative to a surround." After showing us how such a view might be made plausible, Tye nonetheless rejects it: "We experience the redness of a ripe tomato as not involving anything *away* from the facing surface of the tomato [but] as being a local feature of that surface, just as we do its shape. To take a relational view of color is to repudiate this commonsense fact" (*op. cit.*, p. 153).

Tye's favored response to the simultaneous contrast objection is to maintain that simultaneous color contrast is a visual illusion, like some of the stock illusions of shape. It is, to use Mohan Matthen's phrase, "normal misperception."[5] But is color contrast really like those illusions in which straight lines appear bent? One obvious difference is that we have ways of determining that illusions of shape really are illusions. For example, in the Müller-Lyer illusion one can use a ruler to measure the lengths of the lines in question. By what comparable means can we determine that the color change in simultaneous contrast is apparent rather than real? We can, of course, measure the spectral reflectances of the samples. This will tell us that the juxtaposition has not brought about a change in reflectance, but nobody should have doubted that. It will tell us that there has been no *color* change only if we already know that spectral reflectances are colors, and that is, once again, the point at issue.

The stock procedure for showing that a color has not "really" changed is to mask the region surrounding the colored surface in question with an achromatic material. But all this does is to substitute one surround for another in order to offset chromatic changes. We are still left with changes in the perceived achromatic component of the color. If the mask is darker than the sample, the sample will appear lighter, and if it is lighter than the sample, the sample will appear darker. In addition to these lightness changes, there will be changes in the perceived chroma (colorfulness) of the sample. There are but two possible ways to exhibit a target independently of a surround. One of them is to show it against a background of perfect darkness—the null surround. The other is to envelop the viewer's field of view with a region of color that matches the sample, creating a ganzfeld. Both situations have the distinct disadvantage that the sample in question is experienced as pure film color rather than surface color. Its color constancy drops to zero since the ratio comparison that is essential to the operation of the constancy mechanisms has been eliminated. And in the ganzfeld case, in which contrast is absent, the color disappears completely within a few seconds! These are about as far as one can get from "normal" viewing conditions.

---

[5] See "Biological Functions and Perceptual Content," this JOURNAL, LXXXV, [illegible] (January 1988): 5–27.

Eliminating simultaneous contrast in order to see colors as they "really are" has some other serious disadvantages that Tye attempts to address. Without achromatic contrast, black, white, and gray surfaces cannot be distinguished from each other. One is only aware of the sample's brightness or dimness and is unable to apportion the degree of brightness between the reflectance of the surface and the intensity of the illuminant. Thus, the appearance of blackness requires contrast with a lighter surface. Eighty years ago, Adhémar Gelb showed that if one takes a piece of highly reflective paper illuminated at a constant level, and surrounds it with a region of brighter light, the paper can be made to appear any shade of gray whatever, down to and including a very deep black. Indeed, if the light on the sample and the surrounding light can be adjusted independently, any achromatic paper can be made to appear any shade of white, gray, or black. In Gelb's experimental setup, only the contrast ratio matters.[6]

Tye nevertheless maintains that we can objectively specify a black surface independently of contrast as one that is achromatic and reflects less than ten percent of the incident light. But we should notice that Gelb's experiments show that such a surface will look black only if the scene has an average reflectance that is much greater than ten percent. Take an environment in which the average reflectance is about two percent, turn up the intensity of the illuminant sufficiently, and, provided that the scene has no interreflections, Tye's "objectively" black sample will appear white. On the other end of the scale, take the most reflective object that can be manufactured, hold it up against the light of the daytime sky, and notice how gray it appears. The situation with brown is similar. One can make a sample that appears orange appear brown by giving it a surround that is sufficiently lighter. Either the surround can be lightened, or the orange sample can be darkened. Once again, only the ratio between the two matters. And contrary to what Tye asserts, there is no such thing as a brown ganzfeld, and every object that we call 'brown' will look either yellow or orange if the lightness ratio with its surroundings is appropriately altered. This is why there are no (isolated) brown lights, as Ludwig Wittgenstein famously remarked.

Black and brown are, *pace* Tye, pure contrast colors, so to treat all contrast colors as illusory is to eliminate two whole color categories. But worse is yet to come. Even in daylight, *all* light and dark colors depend upon simultaneous contrast if they are to be seen. Take a photograph of a scene with a wide color gamut. Separate the achromatic portion of the photograph from the chromatic portion. The chromatic colors will then be equiluminant. Not only will the chromatic portion fail to contain brown, neither will it contain navy blue, nor olive green, nor any dark, nor any light colors. Remove achromatic contrast, and you remove most of the colors we experience in a typical scene!

---

[6] There are other factors too—see A.L. Gilchrist et al , "An Anchoring Theory of Lightness Perception," *Psychological Review*, CVI (1999): 795–834. Available on-line at http://psychology.rutgers.edu/Users/alan/index.html.

Perhaps Tye can escape from these difficulties by adopting the course of action that he has rejected: let the color of a surface depend not only on its own spectral reflectance, but also upon the reflectances of surrounding elements in the visual scene, suitably weighted by cone sensitivity functions. This tactic has been employed with some success in various computational theories of color vision. It will enable one to deal with the problems of accounting for brown and black that we have just touched upon, and will restore the gamut of colors that Tye's own view denies us. Nonetheless, it falls well short of accounting for the relevant phenomena. Purely perceptual factors that seem to have nothing to do with the physical characteristics of the scene are important determinants of the strength of simultaneous contrast. For instance, whether adjacent colors in an array are seen to contrast or to assimilate depends upon the angular subtense of the elements of the array at the observer's eye rather than upon their physical size. Thus the colors that one sees in a divisionist (or pointillist) painting by Seurat critically depend upon the distance at which the painting is viewed. The perceived geometry of the array matters as well: two surfaces not perceived as coplanar exhibit significantly less simultaneous contrast.[7] And the strength of the effect depends upon whether the center is perceived as figure and the surround as ground, as well as whether the color is seen as surface color or as the color of the illuminant. It is quite unclear how such facts could be explained solely in terms of relations among external stimuli. In fact, no visual scientist doubts that if we are to explain why it is that the color appearance of a surface is a function of the perceived characteristics of its surround, we must look to mechanisms within the brain of the viewer.[8]

---

[7] Gilchrist, "Perceived Lightness Depends on Perceived Spatial Arrangement," *Science*, New Series, cxcv, 4274 (January 14, 1977): 185–87.

[8] See the thorough treatment of the subject in Paul Whittle, "Contrast Colours," in R. Mausfeld and D. Heyer, eds., *Colour Perception From Light to Object* (New York: Oxford, forthcoming).

The last major objection with which Tye deals comes not from card-carrying visual scientists but from a fellow-traveling philosopher. C. L. Hardin[9] argues that the unique-binary structure of the hues is essential to them and is not properly captured by any reflectance-based theory of colors. Tye's response proposes that the unique-binary phenomenal hue structure that is presumably realized by human neural machinery be mapped into the spectral stimuli for color experience. Thus, any spectral stimulus that activates the RED opponent channel and inhibits the GREEN channel but leaves the YELLOW and BLUE opponent channels in equilibrium is to be denominated unique red. Any stimulus that excites the RED and YELLOW opponent channels but inhibits the GREEN and BLUE channels is to be denominated orange, and so on. This is not, of course, a physically natural classification of stimuli, but rather a biologically useful categorization.

Here Tye is employing a Ptolemaic strategy. Instead of reading the phenomenology off of the reflectances of objects, as his doctrine of the transparency of color perception would seem to suggest, he is taking the phenomenology of color categorization and internal color relationships that are structured by the observer's opponent mechanisms and projecting them onto the spectral stimuli. His claim is, of course, that the visual system is selecting visually but not physically significant structures that are latent in the stimuli. On this reading of his projective strategy, colors are like Gibsonian affordances. But as we have just seen, the problem with claiming that the eye-brain merely *selects* colors from the reflectances in the environment is that the actual appearance of those colors is a function of a host of variables that are independent of those spectral surface reflectances. These include the adaptive state of the viewer, the retinal size and location of the light reflected from the target object, the character of the surround, the intensity of the light, and many other factors.

But if we were to put these considerations aside, would Tye's Ptolemaic tactic neutralize Hardin's argument? No, because individuals with normal color vision differ in their chromatic responses. There is, for example, no such thing as *the* equilibrium point for the BLUE/YELLOW channel. Rather, each person has his or her individual equilibrium point, with the consequence that each one has his or her individual spectral locus for unique green. The only way to make Tye's tactic work is to relativize it to each particular observer. It is one thing to let objective color categories be anthropocentric. It is quite another to let them be idiosyncratic.

---

[9] See *Color for Philosophers*, chapter 2.

Tye appears to be unmoved by considerations of observer variability, remarking that this is no more disturbing than to have a variety of somewhat inaccurate measuring sticks. Such a situation, says he, would not give us any reason for supposing that length is not a physical magnitude. Alas, the analogy of variability of measuring sticks to individual color variability is fundamentally flawed. By dint of both theory and a variety of practices we can independently determine the accuracy of any particular instrument that purports to measure length. There is no such independent method for determining the spectral locus of unique green, nor is there any prospect of finding one. Each observer's locus is reliable, but there is no consensus among normal observers. For a color realist, the spectral locus of unique green must be a matter of fact. If it is a matter of fact, there ought to be some way of determining what it is. It is clear enough that the just by looking principle will not suffice, unless we can be given some reason why the looking of some normal observers is to be preferred over the looking of others. The brute fact is that, to date, no color realist has offered even a plausible suggestion of how one might go about making this determination. Indeed, every contemporary visual scientist would regard the search for the "true" locus of unique green to be a fool's errand. For now, at least, Hardin's argument remains untouched, so there is still no good reason for identifying spectral reflectances with the hues that we see.

The importance of individual differences in color perception to the debate about the ontological status of colors has not been fully understood by many philosophers. So let us take a cue from Descartes, and formulate a series of objections and replies:

Objection: Could not the differences in specifying unique green be due to differences in the observers' use of color terms?

Reply: In a carefully controlled test, subjects were asked to estimate the respective percentages of redness and yellowness in an orange test light, or the respective percentages of blueness and redness in a purple test light, and so on. The average results were plotted and compared with the average of the subjects' opponent response functions. The two plots corresponded very closely. The opponent response functions do not depend upon anything more than the ability to turn a knob so as to cancel the chromatic components of test lights presented one at a time. Alternatively, they may be calculated from the subjects' color-matching functions. Neither chromatic cancellation nor color matching require more than minimal linguistic competence for their successful performance. It is safe to say that people's careful, precise, considered use of color terms in controlled situations reflects just what they are seeing, and that disagreements about unique hue loci are not merely disagreements about the use of color terms.

Objection: Should so much weight be placed on small differences in determining just four points in the hue circle?

Reply: The variations in the placement of unique hue loci are considerable. Taken together, the color samples in dispute span at least thirty percent of the total hue circle. The disagreements among observers are real and robust. Wavelength discrimination in the unique green region is exquisitely sensitive, on the order of 2 nm. An observer's unique hue determination is reliable across trials to about plus or minus 2 nm. But the variation in the unique green locus determinations of 50 color-normal observers was 30 nm![10] Moreover, similar disagreements among observers are to be found in a variety of color-matching procedures. One instance is particularly interesting.[11] Using an instrument called the Nagel anomaloscope, a standard instrument for diagnosing color deficiencies, color-normal observers are asked to match a yellow test hemifield with a mixture hemifield of red and green primaries in which the observer can set the red/green ratio. For men, the distribution of ratios is bimodal, falling into two distinct groups, with 60 percent of the observers in one group and 40 percent in the other. The distribution of ratios for women is unimodal, and broader than that for men. In the last decade it has been shown that these distributions are correlated with genetically based polymorphisms of longwave cone photopigments. Here we have a clear case of quantifiable, biologically based individual variations in color perceptions for normal observers under rigorously controlled standard conditions. No scientific sense can be attached to the claim that some of the observers are perceiving the color of the stimulus correctly and others not.

Objection: Suppose we hold with Tye that there are the determinables reddish, yellowish, greenish, and bluish, and that something is unique green just in case it is greenish and not bluish or yellowish. Differences in what are perceived as unique hues will then be differences in what are perceived as determinates of these determinables. This makes the predicates 'reddish', and so forth, vague—there will be clear cases of

[10] Other studies have found a variation of 80 nm for unique green. For a full discussion of the topic, see M. A. Webster and S. M. Webster, "Variations in Normal Color Vision, III: Unique Hues in Indian and United States Observers," *Journal of the Optical Society of America A*, xix, 10 (October 2000): 1951–62.

[11] M. Neitz and J. Neitz, "Molecular Genetics and the Biological Basis of Color Vision," in W. G. K. Backhaus, R. Kliegel, and J. Werner, eds., *Color Vision Perspectives from Different Disciplines* (New York: de Gruyter, 1998), pp. 100–19

their application, but there will also be cases where people will differ, and no way of adjudicating the differences. But the vagueness of color categories should not count against the reality of colors.

Reply: Let us take a vague predicate with which we are well acquainted: 'bald'. I suppose that most people would regard me as instantiating this property, but in a room replete with skinheads I would not be taken to be a good exemplar. In general, there are plenty of cases in which I would be uncertain about whether or not to apply this term. If forced to attach labels in instances such as this, I would expect to be inconsistent in my applications, and could easily be persuaded to shift my criteria. This uncertainty and ready criterion-shifting does not apply to my choice of unique green. Given six Munsell color chips with dominant wavelengths spanning the 490–520 nm range, I, and virtually everyone else whom I have observed making the choice, might vacillate between two adjacent chips, but for any single observer the vacillation rarely if ever extends across the entire range. And yet in most largish groups, each chip will be chosen by at least one person. We are, I think, too ready to assent uncritically to the proposition that all color terms have irremediably vague boundaries. Of course we have trouble answering the question, "Where does red stop and yellow begin?" or "Where does green stop and blue begin?" But now ask this question: "Where does yellow stop and blue begin?" The answer to that is not so problematic, since although something can be both reddish and yellowish, nothing can be both yellowish and bluish. We are not asked to judge the middle of a range, but rather its most extreme limits: a far easier task.

Despite valiant attempts, Tye, like other reflectance theorists, is caught between two incompatible demands of common-sense realism. One is that colors should be features of the surfaces of objects that are independent of human perception. Spectral reflectances satisfy this demand. The other requirement is that colors are normally what we experience them to be. This requirement, we have seen, is not met by spectral reflectances. In particular, the phenomenology of color, including the relations that colors bear to each other, is not well modeled by features and relationships among spectral reflectances. Take away the neural machinery of visual perceivers, and you take away the phenomenal and thus the qualitative characteristics of red, yellow, blue, green, black, and white. But are not these the core subject matter of a theory of color?

Color realists hold that the world contains both spectral reflectances and experiences of color. Color antirealists agree, and point out that spectral power distributions and color experiences are jointly sufficient to explain the gamut of chromatic phenomena. We need not invoke the colors of common-sense realism at all. So why should we include them in our ontology?

C. L. HARDIN

Syracuse University

AGAINST MUSICAL ONTOLOGY*

Many recent philosophers of music have thought it interesting, and possibly even useful, to try to say what sort of thing a musical work *is*, to specify, for example, the conditions that a performance must satisfy if it is to count as a performance *of* the work it purports to instantiate. They have thought it self-evidently worthwhile, in other words, to raise questions about musical ontology; and they have thought that answers to those questions might facilitate, or even be required for, reflection on apparently adjacent areas of interest, such as musical performance. It is my contention here that all such thoughts are mistaken, and that a serious philosophical engagement with music is orthogonal to, and may well in fact be impeded by, the pursuit of ontological issues, and, in particular, that any attempt to specify the conditions of a work's identity must, from the perspective of musical aesthetics, be absolutely worthless.

I

Philosophical enthusiasm for musical ontology was fueled, if not actually created, by the publication of Nelson Goodman's *Languages of Art*,[1] where it was claimed that a musical work is identical to the score in which it is notated, and so that the score "defines" the work (*ibid.*, p. 178). It followed from this, on Goodman's view, that "complete compliance with the score is the only requirement for a genuine instance of a work"—or, to put it rather more starkly, that even "the most miserable performance without actual mistakes [counts] as such an instance, while the most brilliant performance with one wrong note does not" (*ibid.*, p. 186). This, not surprisingly, has struck most readers as wildly implausible—as Goodman, of course, knew that it would. "Could we not," he asks, "bring our theoretical vocabulary into better alignment with common practice and common sense by allowing some limited degree of deviation in performances admitted as instances of a work?" But the answer, he holds, is no, we could not. For

---

*I am grateful to Maria Alvarez, Denis McManus, Alex Neill, and David Owen for their comments on earlier versions of this essay.

[1] Indianapolis: Hackett, 1976. Goodman was not of course the first to discuss musical ontology or the relation between works and performances, but his conclusions certainly heated the discussion up.

0022-362X/03/0004/203–20

> this is one of those cases where ordinary usage gets us quickly into trouble. The innocent-seeming principle that performances differing by just one note are instances of the same work risks the consequence—in view of the transitivity of identity—that all performances whatsoever are of the same work.... Thus while a score may leave unspecified many features of a performance, and allow for considerable variation in others within prescribed limits, full compliance with the specifications given is categorically required (*ibid.*, pp. 186–87).

No one, so far as I am aware, has thought that Goodman is right about this. The general feeling has been that any theory *this* seriously at odds with the intuitions shared by more or less everyone who has ever listened to or played a piece of music must be (and whatever "the transitivity of identity" might suggest) mistaken. It should be noted, however, that Goodman is driven to conclude as he does by reasons that are, in the end, peculiar to his own highly distinctive theoretical priorities—in this case, most proximately, by a markedly rarefied and abstract conception of what it is for something to count as a "notational system" (*ibid.*, pp. 148–57). It is his thinking about "notational systems," that is, that both leads him to the views about works and performances that he espouses, and leaves him relaxed in the face of their extremely counterintuitive character. This suggests that the natural move, if one disagrees with his conclusions, must be to declare oneself motivated by different priorities—by musical ones, say—and pass by on the other side; and I think that this is in fact the right move to make. There is no need, it seems to me, to be drawn into argument here.

That is not, however, how things have gone. What the literature shows instead is a quite remarkable willingness to take Goodman as having thrown down the gauntlet. What, it is asked, if one disagrees with Goodman's conclusions, *is* a musical work identical to (if not the score in which it is notated)? What *is* it for a performance to be of the work it purports to be of (if not getting all the notes right)? Thinking about musical performance has been assimilated, that is, to asking what might constitute the essence of a musical work, and what conditions anything that is to be taken as an "instance" of a given work—for example, a performance of it—must satisfy. Ontological questions, in other words, have been placed firmly in the front line.

For instance, and to tease out just one strand in the post-Goodman literature, a number of broadly Platonist positions have been advanced, in conscious opposition to Goodman's own nominalism. Peter Kivy,[2] for example, has argued for a fairly full-blooded, if forgivably broad-brush, version of Platonism, suggesting that "works are universals, or types, or kinds, performances related to them as particulars, tokens, or instances" (*ibid.*, p. 75); that a "sound structure" (considered as a universal, type, or kind) is "a concrete identity criterion" for a work's being the work that it is;[3] and that such sound structures are notated in scores, where a score, "in a loose sense of 'uniquely',"

---

[2] "Orchestrating Platonism," in his *The Fine Art of Repetition* (New York: Cambridge, 1993), p. 75. For a harder-nosed Platonism, see Nicholas Wolterstorff, *Works and Worlds of Art* (New York: Oxford, 1980).

> uniquely determines a correct performance, under a given set of implicit conventions for interpreting the score—conventions which may be quite different in different historical contexts—and a performance uniquely determines a score under a similar set of historically bound conventions.[4]

Kivy's Platonism, then, identifies a musical work with an (abstract) "sound structure" and regards as "correct" any performance that allows that structure—or something "loosely" like it—to be heard. He also expressly embraces the consequence of his Platonism, namely, that if "works are universals, or types, or kinds" they can neither come into nor go out of existence, and hence that the "composing" of works, so construed, must be a matter of discovery rather than creation (*ibid.*, pp. 38–47). It is partly out of a reluctance to accept this sort of conclusion that Jerrold Levinson[5] has proposed a different kind of Platonism.[6] According to this, a work is to be regarded, not merely as a "sound structure," that is, as "a sequence of sounds qualitatively defined," but as "a compound...of a sound structure and a performing-means structure," where the latter is "a parallel sequence of performing means specified for realizing the sounds at each point" (*ibid.*, p. 78). Levinson suggests that we call the resultant compound "an S/PM structure, for short." S/PM structures, described by Levinson as "implicit types," include every logically possible combination of sounds in every logically possible combination with every logically possible combination of performing means. Such types are Platonic through and through: they can neither come into nor go out of existence. But not all S/PM structures are works; the overwhelming majority of the possibilities that they represent never have been, and never will be, realized. What distinguishes works from other S/PM structures is that they are what Levinson calls "initiated types"—that is, types that "begin to exist only when they are initiated by an intentional human act": works, then, as "initiated types," can be said to be brought into being, created rather than discovered, by the intentional act of composing (*ibid.*, pp. 80–81). A performance of a given work is then specified as "a sound event that is *intended* to instantiate" the work, which "represents an attempt to exemplify" the work's "S/PM structure in accordance with" the composer's "initiation" of it—"and which *succeeds to a reasonable degree*" (*ibid.*, p. 86).

---

[3] Kivy, "Platonism in Music: Another Kind of Defense," in *The Fine Art of Repetition*, p. 63.

[4] Kivy, "Platonism in Music: A Kind of Defense," in *The Fine Art of Repetition*, p. 56 Kivy's sense of 'uniquely' has to be a "loose" one, otherwise his characterization of the relation between score and performance would be identical to Goodman's characterization of the relation between a score, understood as "a character in a notational system," and its "compliance-class," a class which includes performances—see Goodman, pp. 177–78. Kivy needs to keep his distance from Goodman, not only because he rejects Goodman's nominalism, but because he wants his own conclusions to be intuitively plausible.

[5] "What a Musical Work Is," in his *Music, Art, and Metaphysics* (Ithaca: Cornell, 1990), pp. 63–88.

[6] Or, more strictly, had already proposed a different kind. Levinson's work on the topic pre-dates Kivy's by several years.

These examples should give some indication of the sorts of effort that have been made to address the ontological questions I mentioned a moment ago, while also attempting to avoid some of the counterintuitive consequences of Goodman's position. There is, so far as I am aware, no consensus—nor even a hint of it—on these matters, except perhaps for a growing feeling that the sheer diversity of things that might, in various times and places, be thought of as music or as performance makes it unlikely that any monolithic account, however tolerant of deviation, will suffice. A more catholic approach to ontology has emerged accordingly—an approach perhaps most comprehensively evinced by Stephen Davies's recent book, *Musical Works and Performances: A Philosophical Exploration*.[7] This does, as I say, represent a development, but it does not represent any kind of departure: ontological questions still dominate.

My view, as I said at the outset, is that the whole move to ontology in thinking about music is a mistake. Before saying why, though, I should make clear how I will be taking the term 'performance' from now on. By 'performance' I will mean not only the playing of a work by an individual or group before an audience, but also recordings, transcriptions, arrangements, versions and, in general, renditions of every sort. I take my lead in this from Kivy,[8] who argues, first, that a performance is "an arrangement or version of the musical work that has been performed, and, as such, a subject of the kind of evaluation and aesthetic satisfaction that artworks support and provide," and, second, that the "performer is an artist, somewhat akin to a composer or, better, 'arranger' of musical works" (*ibid.*, p. 261). The relatively capacious conception of performance that results will allow me, I hope, to make the argument I want to make in a reasonably direct manner, without getting bogged down in (or indeed traducing) such finer-grained details as might or might not be at issue.

---

[7] New York: Oxford, 2001.

[8] *Authenticities: Philosophical Reflections on Musical Performance* (Ithaca: Cornell, 1995).

II

When was the last time you came away from a performance of a piece of music—live or recorded—seriously wondering whether the performance had been of *it*? My guess is, never. Even if you were to hear the Bach-Busoni *Chaconne*, say, as played by me, or by some comparably giftless pianist, it still would not occur to you to doubt that the victim of the musical murder you had witnessed was, indeed, the *Chaconne*. In fact, if it were played any worse, in such a way that the victim was actually unidentifiable, there would be no reason to think that you had witnessed any sort of performance at all—no reason to think, that is, that you had witnessed a performance *of* anything. Indeed, it is precisely because doubts of the relevant sort have next to no tendency to arise that muzak'd versions of, say, the *Ode to Joy* theme have the power to enrage and depress that they do. One would not suffer as one does in elevators and supermarkets if the doubts were real. One would not be reduced to misanthropic cursings, to mutterings of "How *could* they? How *dare* they?" if it really did strike one as a serious possibility that the miserable, denatured pap oozing from the speakers was not the Beethoven after all. It clearly *is* the Beethoven, and that is why it makes one feel so low and vicious. In cases like these, then, one's doubts are not about what a rendition is *of*—far from it—but about whether any penalty could feasibly reflect the gravity of the offense. These thoughts, admittedly much understated, suggest two conclusions. The first is that, in our ordinary—indeed in our actual—aesthetic encounters with renderings of pieces of music, our primary concern, or at the very least one of our most prominent concerns, is whether a given rendition is any good, or, if it is not, whether it is so bad as to merit further action. As listeners, that is, we are chiefly alert and sensitive to issues about the value of what we are hearing. The second conclusion is that, since these sensitivities are operative, and even virulent, against the background of an apparently rather robust sense of work identity, issues concerning work identity can hardly be very urgent if what we are chiefly interested in is our aesthetic experience of renditions of pieces of music. If we are doing aesthetics, that is, ontological questions deserve a place in the back row, at best.

I think that both of these conclusions are correct. Neither, however, commands evidently overwhelming support from contemporary philosophers of music, as a moment's reflection on the material mentioned in section I will confirm. What we find there is, essentially, a range of views about the room for maneuvering that a performance has before it ceases to be a performance of what it was supposed to be of, backed up by a range of views about what constitutes a work's essence or identity. From this sort of starting point, it is hardly surprising that evaluative issues, where they get a look in at all, are confined to questions of legitimacy—to asking whether this or that performance of a given work is faithful enough to whatever happens to be the favored set of identity conditions to count as acceptable, legitimate instances of performances *of* it. The question whether this or that performance, or style of performance, is actually any good, or is minimally worth listening to, is scarcely raised.[9]

If one is serious about the philosophy of music, this last fact should strike one as scandalous. Not as inexplicable, perhaps: philosophers are drawn to metaphysics, and the chance to go on about identity conditions, and related ontological matters, has possibly been found too good to pass up. But one really ought to be shocked, even so. The most direct way to bring out why, perhaps, is to note that an indifference to genuinely evaluative issues—to issues that animate the experience of actual listeners when they listen to actual pieces of music—presupposes a sharp distinction between what it is to take a philosophical interest in music and what it is to take a critical interest in it. It is true that such a distinction can be drawn. It is true, that is, that the philosophy of music is not identical to music criticism. But the distinction is not, and cannot be made to be, a sharp one, for unless one's philosophical engagement with music is driven by, and is of a sort that might pay dividends for, one's musical experience—including one's evaluative experience—there is no obvious sense in which one is engaged in philosophical *aesthetics* at all. At best, and instead, one might be engaged in the metaphysics of music, and that is a very different activity.

I can imagine two kinds of response to what I have just said. The first is simply to grant that what is being done is indeed musical metaphysics rather than musical aesthetics: If one has made no attempt to do aesthetics, after all, how much of a scandal can it be that one's results are aesthetically inert? I propose, for the moment, to accept this response—although I will be coming back to it. It is the second response, however, that most self-avowed philosophers of music are likely to reach for. According to this, one cannot hope to do justice to the kinds of evaluative concern that I have mentioned unless one has got one's metaphysics—one's ontology—into good shape first. There is no point, that is, in asking whether a performance of such-and-such is of any value until one has determined that that performance is, indeed, *of* such-and-such; and this means that one must first work out what it *is* to be such-and-such, that is, determine the relevant identity conditions, and then satisfy oneself that the putative performance meets them. Only after that has been done, the response concludes, can one ask whether the performance is, aesthetically, any good—first things first.

---

[9] There are exceptions, of course—the most honorable recent one, in my view, being the raising of just that question by Kivy in his generally splendid *Authenticities*, see pp. 155–61.

This way of thinking strikes me as quite profoundly mistaken; and the examples I have already given, the muzak example especially, show why. Imagine yourself seized with a furious sadness at some particularly vile, beat-enhanced, edges-smoothed-over version of the *Ode to Joy* theme. And now imagine yourself thinking as the second response invites you to. On the face of it, things could go in either of two ways, depending on whether one takes one's ontology prescriptively or descriptively. Taken prescriptively, one might decide that what one was hearing did not meet one's own rather stringent conditions for work identity, that is, that the muzak was not really, and despite appearances, a rendition of Beethoven's theme at all. But in which case, why the rage? Or—and still prescriptively—one might decide that the muzak did meet one's conditions, in which case it is hard to see how that decision, given one's enraged response, has left one any the wiser. And taken descriptively the result is the same: one's enraged response shows either that one's identity conditions are wrong (if the muzak does not satisfy them) or that they are superfluous (if it does)—superfluous in the sense that their deployment is not any sort of precondition of pointfully raising questions of aesthetic value. This way of thinking is surely a matter of putting misleading or redundant things first, not first things.

It will be objected, no doubt, that I have stacked the deck in my favor by stipulating that what one hears fills one with rage and depression. It is only this stipulation, after all, that makes preemptive ontological beavering look either misleading or redundant. But this objection has force, clearly enough, only if ontological considerations, in musical contexts, are allowed to trump one's experience of what one actually hears. And the single case in which that possibility has any potential bite is the first one I considered—where one decides, on prescriptive work identity grounds, that what one is hearing is not the *Ode to*

Joy theme at all, and therefore, presumably, that one should not (necessarily) mind it. But here, I have to say, I begin to lose touch with what might be being said. In the sort of example I am imagining—which is to say, a case where the *Ode to Joy* theme is clearly being murdered—one could only, surely, decide that it was not that theme after all if it had first occurred to one that it might be, if it sounded enough like it to be worth giving it the ontological once-over for that possibility, specifically. But now it looks as if the decision to deny the title *Ode to Joy* to what one is hearing (assuming that that is the decision one arrives at) cannot be anything more than an elliptical and rather unhelpful way of expressing one's disapproval at what has been done to Beethoven's original. The idea of wondering, perfectly seriously, and from scratch, as it were, whether the prevailing ontological conditions are such as to warrant one's disapproval is either impossible to understand, it seems to me, or else a symptom of something close to aesthetic autism.

But perhaps I am tilting at straw men here. Perhaps the position I should be considering is not, as I have been assuming, that one must always, in every case, get one's ontology straight before evaluative questions can pointfully be raised, but rather that evaluative questions will, in general, be more perspicuously framed if raised against the backdrop of a convincing ontology. And if this is right, it might be thought, metaphysically inclined philosophers of music need neither be flummoxed in the face of their muzak-inspired paroxysms, supposing they have them, nor fearful for their status as aestheticians.

On the face of it, there is no doubt that this modified and more modest position is an improvement on its predecessor. But it is still deeply unpersuasive. How, exactly, is a convincing ontological backdrop supposed to lend perspicuity to evaluative questions? No one, so far as I am aware, has actually asked this: certainly no one has given any sort of explicit answer. But we can begin to see how such an answer might go from something that Davies says:

> To be of a work, a performance must satisfy three conditions. There must be a suitable degree of matching between the performance and the work's contents, the performers must intend to follow most of the instructions specifying the work in question..., and there must be a robust causal chain from the performance to the work's creation, so that the matching achieved is systematically responsive to the composer's work-determinative decisions (*op. cit.*, p. 5).

Davies is here addressing the metaphysical-cum-evaluative question—what conditions must a performance of a work satisfy if it is to count as a legitimate performance *of* it? And his answer—that the performance must "match" the contents of the work, where those contents are determined by the composer's decisions as embodied in the instructions (as found, for instance, in the score) which mark out the work as the work that it is—gives us what I have been calling the "ontological backdrop." According to this, then, a performance is legitimate if it "matches" the work's contents. How might one attempt to understand this backdrop as lending perspicuity to the framing of genuinely evaluative questions? What, in other words, if this is the backdrop, does it suggest that a question about the value of some particular performance ought to look like? The only possible answer is this: How *well*, how *closely* does the performance "match" the contents of the work?—with the presumption that the better it "matches" the better it is. One can see how someone might offer this sort of formulation in an effort to capture what we mean when we commend a performance for its faithfulness to the work performed; and indeed, one can see how someone might think that, in order to make or to understand such a commendation, something appropriately ontological must be going on in the background. But a moment's reflection should be sufficient to show that the notion of "matching" is incapable of doing any such work.

The problem, of course, is that to make the value of a performance a function of how well it "matches" the contents of the work performed, it would be necessary to be able to specify, quite independently of this or that performance, what, precisely, those contents were—what exactly, in other words, the performance was supposed to "match." It would be necessary, that is, to be able to give an exhaustive prescription for the production of a performance that was, not merely legitimate, but excellent, first-rate, admirable. And that is, of course, quite impossible, in just the same way and for just the same reasons that it is impossible to give an exhaustive prescription for the production of a great work of art. However apt Davies's talk of "matching" might be in the context of *legitimacy*, then, it certainly fails to carry over into genuinely evaluative contexts. His ontological reflections, that is, do nothing—as backdrop—to assist in the perspicuous framing of evaluative questions. Indeed, quite the reverse: they actively encourage the framing of those questions in a misleading way. Nor is this a problem peculiar to Davies. I have chosen to discuss Davies explicitly only because he, explicitly, refers to "matching." But the point is quite general: to the extent that one's ontological reflections are geared to the specification of work identity conditions—that is, to conditions that might be satisfied by a *range* of different performances—one is necessarily committed to the specification of conditions that are, in the operative sense, *not* specific to any performance in particular—that

Things get even worse for the purveyor of identity conditions, as we can see if we ask what *is* meant by a performance's being "faithful" to a work—the genuinely evaluative notion, as I have just suggested, that no talk of "matching" can hope to capture. Much might be said about this, and a full treatment of the question would take us far afield. But it is sufficient for present purposes to note that, whatever the details of a full account should turn out to be, this much, at least, would have to be included among them—that a performance of a work cannot be "faithful" to it unless it evinces an understanding of it. And if a performance's faithfulness is, minimally, a matter of the understanding it shows, then a performance is, in that much, to be valued in proportion to the richness, depth, insight, subtlety and so on of the understanding it evinces.[12] But if this is right, evidently enough, much of the "content" of a given work is only revealed *in* the understandings that faithful performances of it evince. And that means that any attempt to specify that content—the content to which a good performance is faithful—*in advance* of evaluative judgments about particular performances of it, or independently of such judgments, must be futile and self-defeating. The fact is, in other words, that since faithfulness is a matter of understanding, and since understanding is not a matter of bringing off a "match" of any sort with some independently specifiable content, nothing *at all* can be said in advance of any particular faithful performance about what "faithfulness," in its case, will amount to. This is why some performances, in evincing an original or especially penetrating understanding of the work performed, and so in bringing to light aspects or dimensions of its content that other performances have not, are described as "revelatory."

But why is this bad news for the ontologist? Or why, rather, is it *more* bad news, rather than just a gloating restatement of the bad news already announced, namely, that the specification of identity conditions cannot hope to do any evaluative work? Here is why. Until just now, it might have seemed that the evaluative impotence of identity conditions at least left in place their claim to *be* identity conditions—that is, their claim to be the arbiters of legitimacy, even if not of quality. It might have seemed, in other words, as if aspects of a work's content not revealed exclusively through performance might be specified independently, and taken, if a performance

---

[12] The notion of "understanding" cannot, of course, be reduced to the notion of "matching," for standard Wittgensteinian reasons—see *Philosophical Investigations*, §§527–37

"matches" them, to show that the performance is indeed a legitimate performance *of* the work. But even this toehold is now denied. For if, as I have argued, a revelatory performance is one that is faithful to a work in a way that necessarily defies specification in advance, then, first, as a performance that is faithful to the work it is, trivially, *of* it, and second, amongst its unforeseeable qualities may well be the disregard of any, or even of all, of the independently specifiable bits of "content" that it was supposed, as a legitimate performance, to have to match. This is not to deny that such bits of "content" may indeed be specifiable independently. But it is to deny that their being so specifiable accords to them any special authority. It is just as futile, in other words, to hope that one's identity conditions will trump the evidence of one's ears in this case, where the performance strikes one as masterly, as tremendous, as it is in the case discussed earlier, where the performance strikes one as an infuriating, depressing travesty. And this means that one's "identity conditions," whichever ones they happen to be, cannot be *identity* conditions, or even conditions. If an exemplary performance can violate them, then they are, at best, an expression of some more or less reasonable set of expectations, or, as one might rather more grandly put it, a set of defeasible criteria. And this means, in turn, that they lose any claim to report the fundamental ontology of the matter: if some—or perhaps even all—of a work's allegedly identity-confirming properties can fail to be reflected in a faithful performance *of* it, then those properties can have nothing to do with the work's being what, essentially, it is. Rather, if what I have been arguing is right, one finds out what a work is, what properties it has, by experiencing performances of it, or by giving performances of it, and that is a process of discovery that may well have no determinate end.

The upshot of all this is not only that neither of the possible ways of cashing out the second response to the charge of "scandal" is able to deliver, but that retreat to the first response—to the claim that what ontologically inclined philosophers of music are doing is musical metaphysics, not musical aesthetics—is doomed to failure too. For what the foregoing shows is that ontological questions about pieces of music are only perspicuously to be framed, if they are to be framed at all, against an *aesthetic* backdrop of *already-answered* questions about the value of performances *of* them. Strawson is again helpful here:

> To use a fashionable phrase, the *criterion of identity* of a work of art is the totality of features which are relevant to its aesthetic appraisal.. . Perhaps I could also express the point in this way: the only method of describing a work of art which is both...adequate for the purpose of

> aesthetic appraisal, and does not use evaluative language, is to say, 'It goes like this'—and then reproduce it. And, of course, this is not a method of *describing* at all (*op. cit.*, p. 185).[13]

If the second response to the charge of "scandal" put the ontological cart before the evaluative horse, and so forfeited its claim to be engaged in either ontology or aesthetics, although not its claim to be engaged with music, albeit in a thin way, this first response, in attempting to put the horse and cart in different fields entirely, forfeits any claim to be concerned with *music*, with a "method of *describing*" music, at all. Its proponents must cut themselves off from what Davies calls a work's "content"—that is, from what can be understood and appreciated in it. And in doing so they confine themselves, at best, to the metaphysics of sounds that might, conceivably, be heard as music, or maybe to the metaphysics of scores. Musical metaphysics is predicated on musical aesthetics; and in musical aesthetics, ontology comes last (at the end of time, perhaps).

III

One might wonder, if the argument of the previous section is right, why it is that the temptation to begin thinking about performance in terms of ontology and identity conditions should have been felt at all. In part, and I am convinced of this, it does have to do with the lure of metaphysics: metaphysics is grown-up philosophy, and insecure aestheticians may reach for it too readily. It also has to do with the unwittingly baneful influence of Goodman, as I suggested at the beginning of this essay. But I can also think of two less ad- or ab-hominem sorts of reason, both superficially understandable, but neither, when one looks at all closely, compelling. The first concerns "hard cases"; the second, first performances. I will take them in turn.

"Hard cases": the examples I have used have tended to suppress the possibility of genuine borderline cases, cases in which it really is not clear whether a given "performance" is of a given work. It is all very well to concentrate on obvious outbreaks of musical murder, but sometimes, at least, things are not so straightforward. What is one to make, for instance, of a case in which one is genuinely puzzled by what one is hearing—not enraged or bilious, just puzzled? It is *possible*, one thinks, that one is hearing a performance of such-and-such. But, then again, one might not be. Surely, it will be said, this sort of case, a genuine borderline case, shows precisely why one needs to have some ontological views to draw upon, some identity conditions that will, at least in principle, allow one either to settle the question, or else to characterize the borderline on which the case sits in a perspicuous way. And surely neither result would be aesthetically worthless: ontology cannot, then, be written off so quickly.

---

[13] I take it that Strawson's "totality" of relevant features is exactly what the process of discovery I refer to would, at some notional end point, have brought to light.

I am perfectly happy to agree, at least for the sake of argument, that there might be genuine borderline cases. I am ready to agree that it might not always be clear, or may be simply undecidable, whether a particular "performance" is, indeed, of a given work. But this fact, if it is one, does not have any of the redemptive consequences for ontology envisaged. For one thing, and in light of the argument of the previous section, it would be a bit odd, to put it no higher, if ontology suddenly turned out to be splendid for borderline cases while being irrelevant or misleading for uncontentious ones. But be that as it may, the problem with the present suggestion is that it simply *ignores* the argument I have offered. Suppose that one were able to decide, in some particular borderline case, that a performance was, in fact, of such-and-such a work. This would indeed be because one had decided that it satisfied certain conditions—conditions relating, presumably, to the "content" of the work in question. But since, as I have argued, and I think shown, that content is crucially revealed by the understandings evinced in faithful performances of the work, one's capacity to decide in favor, as it were, of some particular putative performance represents a triumph, not for ontology, but for aesthetics. And the same goes if what is sought is a perspicuous characterization of the borderline that a "performance" occupies: that is going to be yielded, not, say, by a list of the identity conditions that the "performance" matches or fails to match, but by evaluatively driven critical inquiry. As far as I can see, then, "hard cases" offer no comfort at all to those who would like to see ontology done first. At most, it needs to be done last.

First performances: according to me, much of the "content" of a given work is revealed through the understandings evinced of it by faithful performances. But what if there have *been* no performances, faithful or otherwise? What if *this* performance is the first one? There is now no "aesthetic backdrop," as I put it a moment ago, against which ontological questions might be raised. And yet there is surely an answer to the question: Is this (first) performance a performance of the work it purports to be of? And if there *is* an answer to that question, that surely shows that one can determine the identity of a work on independently specifiable grounds—on the grounds, that is, of precisely the sorts of ontological consideration that I have been so busy maligning. So again, the suggestion goes, ontology is not to be seen off so quickly.

My response to this point begins with what is, essentially, an expression of puzzlement. What, exactly, is the difficulty supposed to be here? One shows up at a concert hall expecting to hear, as the program promises, the first performance of such-and-such by so-and-so; the orchestra have their scores open in front of them, the conductor does the usual conductorly things, and they play a piece one has never heard before. How, precisely, is the question whether this (first) performance is a performance of the work it purports to be of meant to arise? To what dark suspicions would one have to be prey for this to strike one, genuinely, as a pressing or pointful question? What I am trying to suggest, in other words, is that if all of the surrounding circumstances are exactly as one would expect, if everything is outwardly consistent with the performance's being what it purports to be, namely, the first performance of such-and-such, the question whether it *is* of what it purports to be of simply does not come up. The context does not invite the question, and the motivation for asking it is, as things stand, wholly inscrutable. Skepticism for its own sake is as difficult to make sense of here, in other words, as it is anywhere else. For the question to be worth asking it must, first, be raised in acknowledgment of ordinary contextual considerations and, second, be motivated by factors that might, given those considerations, be intelligible grounds for doubt. So there really do have to be some dark suspicions in play—some genuine reasons to wonder whether what one is hearing really is what it was said to be—before the question, "Is the performance *of* the work?" is so much as askable. Let us imagine some cases.

First, imagine that one is familiar with all of the composer's previous works, and that what one is hearing sounds as if it could not possibly have come from the same hand. This might well be perplexing, and it might well make one wonder what was going on. Possible explanations? Perhaps one is not as good as one had thought at making judgments about composers' styles. Or perhaps the composer's style has changed, or he has simply written something uncharacteristically good or bad. Or perhaps the program has been altered at the last minute. Or perhaps the performance is a hoax, and not of a work by the advertised composer at all. The question we are concerned with can be answered or dealt with in each case. In the first two, and in the absence of further complicating details, the answer is yes, the performance is indeed *of* such-and-such. In the third case, where the program has been changed, the state of affairs originally purported has ceased, the relevant way, to *be* purported, and so the question does not arise. And in the final case, the answer is no, the performance is a hoax—it is not what it presents itself as being. How have these answers been arrived at? Certainly not through the specification of identity conditions. All that is required for any of them is the wholly uncontentious observation that composers standardly compose their own works—an observation that is consistent with any minimally plausible ontology, and so presupposes none.

Second, imagine that one has had a sneak preview of the score, and that what the orchestra is playing seems to have nothing whatever to do with it. Again, one would wonder what was going on. And, again, there might be several explanations. It might be that one is very bad at reading scores. Or it might be that the score one saw was a forgery. Or it might be that the program has been changed at the last minute. Or it might be that the performance is a hoax. The answers to the question whether the performance is of what it purports to be of are exactly the same as in the previous case: yes, yes, not an issue, and no. And this time, all that is required to arrive at them is the humdrum truth that works are standardly notated *in* and performed *from* scores, again not a thought that either presupposes or suggests any particular set of ontological commitments rather than any other.

Finally, and even less interestingly, suppose that one has seen the score in advance, and that one creates for oneself a vividly imagined performance in one's head. And then suppose that the performance one hears in the concert hall is not like that at all. But this is hopeless. First, if the performance one hears is not like that *at all*, this example collapses into the previous one. But if it *is* enough like it to make one concerned about the divergence, then a dilemma presents itself. If, on the one hand, it is granted that the imaginary rendition is equivalent to a genuine performance, then (a) the performance one hears in the concert hall is not, in the relevant sense, a *first* performance, and so cannot prompt the question we are supposed to be worrying about; and (b) the most, anyway, that the divergence at issue could indicate is that the actual performance is either significantly better or significantly worse than the imagined one (and not, that is, that the actual performance was not of what it was meant to be of). Whereas if, on the other hand, the imaginary rendition is denied the status of "performance," its divergence from the actual performance can indicate only (a) that actual performances are different from nonperformances; and (b) that that difference is sufficient to entail the collapse of this example into the previous one, as before. Neither alternative can give the example even the appearance of bite. However one looks at it, it seems to me, first performances pose no threat whatever to my contention that musical ontology is, at best, superfluous if what one is interested in is musical aesthetics.

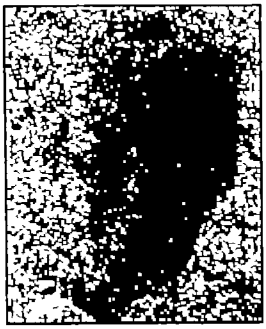
But have I not, it might be asked, committed myself to a whole set of ontological claims throughout the course of my own argument? Have I not claimed, in effect, that works are identical to faithful performances *of* them? And have I not nailed my colors to a firmly realist view of the properties of works that faithful performances reveal? And have I not just helped myself—a bit conveniently—to thoughts about the relations between composers and works and between works and scores that look, for all the world, like covertly specified identity conditions? Have I not, in short, shot myself repeatedly in the foot?

It will come as no surprise to learn that I think the answer is no, I have not. I have nowhere, in effect or otherwise, claimed or assumed that works are identical to (some? all?) faithful performances of them, and nor can I see why anyone would want to. All that I have argued is that performances can show us things about works; and that requires nothing more than the thought that (some) performances are interpretations *of* works—not, I surmise, a proposition likely to provoke a storm of protest, and certainly not one that involves or presupposes (or should prompt) the slightest flicker of ontological reflection. With respect to the second charge, it is true that I have cast a number of my sentences in undeniably realist terms. It is true, that is, that I have spoken of aspects of a work's content as being *revealed* by faithful performances of it, rather than as being brought into existence by them, say. But I have put it like that largely as a matter of convenience, and also out of deference to the convention that, unless there is some good reason not to, one should assume that what one is talking about is, in a straightforward and commonsensical way, real. I do not insist upon this, though. If someone feels more comfortable with the thought that performances "create" content rather than "discover" it, that is fine by me: nothing in what I have said or want to say depends on the distinction. And finally, to the suggestion that I have really just helped myself covertly to a raft of identity conditions, I simply deny that I have. At most, I have helped myself to some perfectly neutral, pre-theoretical thoughts. And these are, first, of a sort that no one, whatever their ontological views, could possibly object to, and second, of a sort which, if they were (by someone else) misconstrued as quasi-theoretical items, would not thereby become identity conditions, but—as I have already argued—expressions, merely, of some more or less reasonable sets of expectations. My position, it seems to me, is steadfastly devoid of ontological commitments; and it is, for the reasons I have given, in much the better shape for that. Musical ontology may, at first blush, appear to be an entirely harmless and even a proper philosophical pursuit. But if it is the philosophy of music that one is interested in doing, then ontology really is an idle distraction, or worse, and I propose that we should have nothing further to do with it.

AARON RIDLEY

University of Southampton

Knowing.



New in paperback

Goodness and Advice

Judith Jarvis Thomson

Edited by Amy Gutmann

How should we live? What do we owe to other people? In *Goodness and Advice*, the eminent philosopher Judith Jarvis Thomson explores how we should go about answering such fundamental questions

"A highly civilized, lively and provocative exchange between interesting people of diverse backgrounds and positions ... A very fine example of a first-rate mind taking great pains to deal carefully, clearly and methodically with perennial problems of the utmost significance."
—Saul Smilansky, *Times Literary Supplement*

The University Center for Human Values Series

Paper $14.95 ISBN 0-691-11473-0

New in paperback

Human Rights as Politics and Idolatry

Michael Ignatieff

Edited and Introduced by Amy Gutmann

With commentary by K. Anthony Appiah, David A. Hollinger, Thomas W. Laqueur, and Diane F. Orentlicher

"This book ... confirms Michael Ignatieff's reputation as one of the most interesting commentators on human rights issues.... An excellent little book."
—David Petrasek, *Ethics and International Affairs*

The University Center for Human Values Series

Paper $12.95 ISBN 0-691-11474-9

Princeton University Press

New from
Oxford

Thinking Without Words
JOSÉ LUIS BERMÚDEZ
Bermúdez offers a conceptual framework for treating human infants and non-human animals as genuine thinkers. The book is written with an interdisciplinary readership in mind and will appeal to philosophers, psychologists, and students of animal behavior
2003 $35.00

A Priori Justification
ALBERT CASULLO
"I expect this book to be the central work in the epistemology of the a priori for years to come. It provides detailed, penetrating, and judicious discussions of the major theses and arguments about the nature, possibility, and objects of a priori knowledge and justification." —Panayot Butchvarov, *University of Iowa*
2003 $55.00

Morality's Progress
Essays on Humans, Other Animals, and the Rest of Nature
DALE JAMIESON
Morality's Progress is the summation of nearly three decades of work by a leading figure in environmental ethics and bioethics The twenty-two papers here are invigoratingly diverse, but together tell a unified story about various aspects of the morality of our relationships to animals and to nature.
2003 paper $24.95 cloth $75.00

New in Paperback!
Virtue, Vice, and Value
THOMAS HURKA
"This book will be the best book on virtue since Aristotle. It will be to virtue what Sidgwick was to utilitarianism "
—Brad Hooker, *University of Reading*
2000 (paper June 2003) paper $24.95 cloth $60.00

Getting Even
Forgiveness and Its Limits
JEFFRIE G MURPHY
"In a voice that is reasonable, incisively witty, finely tuned to human emotion, and wise, Murphy teaches us how to think about our most difficult moral dilemmas When should we forgive? When might it be healthy to hold a grudge? We would all do well to think through these questions from both a personal and moral perspective with this thoughtful and fascinating meditation."—Sharon Lamb, author of *The Secret Lives of Girls* and *The Trouble with Blame*
2003 $21.00

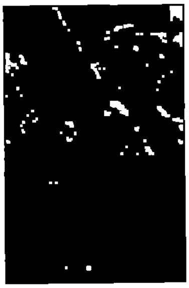
The Oxford Handbook of Practical Ethics
Edited by HUGH LAFOLLETTE
The Oxford Handbook of Practical Ethics is a lively and authoritative guide to current thought about ethical issues in all areas of human activity—personal, medical, sexual, social, political, judicial, and international, from the natural world to the world of business It will prove essential reading for academics and students across a wide range of disciplines
2003 $110.00

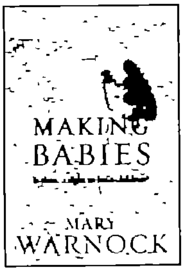
Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup-usa.org

OXFORD

New from

Oxford





Prices are subject to change and apply only in the US To order, or for more information, please call 1-800-451-7556
Visit our website at
www.oup-usa.org

OXFORD

New in Paperback!

A Tear Is an Intellectual Thing
The Meanings of Emotion
JEROME NEU

This book raises fundamental questions about what sustains and threatens our identity. The discussion throughout is informed and motivated by the Spinozist hope that understanding our lives can help change them, can help make us more free.

2000 (paper 2002) paper $24 95 cloth $49 95

Thinking about Consciousness
DAVID PAPINEAU

Papineau argues that consciousness seems mysterious not because of any hidden essence, but only because we think about it in a special way. He exposes the resulting potential for confusion, and shows that much scientific study of consciousness is misconceived.

2002 $29 95

Ethical Intuitionism
Re-evaluations
Edited by PHILIP STRATTON-LAKE

Ethical intuitionists believe that there is an irreducible plurality of basic moral principles, and that rightness and goodness are indefinable properties that cannot be understood in wholly naturalistic terms. In this collection, leading moral philosophers consider how these views are to be understood, and what implications they have for our understanding of morality.

2002 paper $24 95 cloth $70 00

Is There a Right to Have Children?
Making Babies
MARY WARNOCK

Warnock analyzes what it means to claim something as a "right," examines the ethical problems faced by artificial insemination, in-vitro fertilization, and surrogacy, and argues that in the future human cloning may well become a viable and acceptable form of treatment for some types of infertility.

2002 $13 95

New in Paperback!

Boundaries and Allegiances
Problems of Justice and Responsibility in Liberal Thought
SAMUEL SCHEFFLER

"These essays display a remarkable combination of philosophical acuity, moral seriousness and political realism, and deserve to be read closely and repeatedly for the light they shed on our contemporary moral and political predicament."
—David Miller, *Philosophical Books*

2001 (paper 2003) paper $19 95 cloth $29 95

Views into the Chinese Room
New Essays on Searle and Artificial Intelligence
Edited by JOHN PRESTON and MARK BISHOP

Searle argued that, although machines can be devised to respond to input with the same output as would a mind, machines—unlike minds—lack understanding of the symbols they process Nineteen essays by leading scientists and philosophers assess, renew, and respond to this crucial challenge

2002 paper $24 95 cloth $74 00

Best in Philosophy

Aquinas: Political Writings
Edited by R. W. Dyson
Cambridge Texts in the History of Political Thought
0-521-[illegible] Hardback $[illegible].00
0-521-[illegible] Paperback $22.00

Natural Rights and the Right to Choose
Hadley Arkes
0-521-81[illegible] Hardback $[illegible] 00

Augustine: On the Trinity
Edited by Gareth Matthews
Cambridge Texts in the History of Philosophy
0-521-[illegible] Hardback $[illegible].00
0-521-[illegible] Paperback $20.00

Classic and Romantic German Aesthetics
Edited by J. M. Bernstein
Cambridge Texts in the History of Philosophy
0-521-[illegible] Hardback $[illegible].00
0-521-00111-[illegible] Paperback $22.00

The Cambridge Companion to Simone de Beauvoir
Edited by Claudia Card
Cambridge Companions to Philosophy
0-521-79[illegible] Hardback $60.00
0-521-79[illegible] Paperback $22.00

Heidegger's Analytic
Interpretation, Discourse and Authenticity in *Being and Time*
Taylor Carman
Modern European Philosophy
0-521-[illegible] Hardback $60.00

Must We Mean What We Say?
A Book of Essays
Updated Edition
Stanley Cavell
0-521-82[illegible]-6 Hardback $70.00
0-521-52[illegible] Paperback $25.00

Radical Interpretation in Religion
Edited by Nancy K. Frankenberry
0-521-81686-6 Hardback $65.00
0-521-01705-X Paperback $24 00

The Creation of Art
New Essays in Philosophical Aesthetics
Edited by Berys Gaut and Paisley Livingston
0-521-81234-[illegible] Hardback $[illegible].00

Quine and Davidson on Language, Thought and Reality
Hans-Johann Glock
0-521-82180-[illegible] Hardback $[illegible].00

The Cambridge Companion to Darwin
Edited by Jonathan Hodge and Gregory Radick
0-521-77197-[illegible] Hardback $70.00
0-521-77730-5 Paperback $[illegible].00

from Cambridge

Theoretical Philosophy, 1755–1770
Immanuel Kant
Edited by David Walford and Ralf Meerbote
The Cambridge Edition of the Works of Immanuel Kant in Translation
0-521-53170-5 Paperback $35.00

The Philosophy and Politics of Bruno Bauer
Douglas Moggach
Modern European Philosophy
0-521-81977-6 Hardback $60.00

Realism and Christian Faith
God, Grammar, and Meaning
Andrew Moore
0-521-81190-0 Hardback $65.00
0-521-52415-6 Paperback $23.00

More: Utopia
Revised Edition
Edited by George M. Logan and Robert M. Adams
Cambridge Texts in the History of Political Thought
0-521-81925-3 Hardback $30.00
0-521-52540-3 Paperback $10.00

Nietzsche: Writings from the Late Notebooks
Edited by Rüdiger Bittner
Cambridge Texts in the History of Philosophy
0-521-80405-1 Hardback $50.00
0-521-00887-5 Paperback $18.00

Divine Action and Modern Science
Nicholas Saunders
0-521-80156-7 Hardback $60.00
0-521-52416-4 Paperback $22.00

Prices subject to change.

Schleiermacher: Lectures on Philosophical Ethics
Edited by Robert B. Louden
Cambridge Texts in the History of Philosophy
0-521-[illegible] Hardback $60.00
0-521-00767-4 Paperback $21.00

Visions of Politics
3-Volume Set
Quentin Skinner
0-521-81302-4 Hardback $180.00
0-521-89075-6 Paperback $65.00

The Philosophy of Social Practices
A Collective Acceptance View
Raimo Tuomela
0-521-81850-5 Hardback $60.00

Republicanism: A Shared European Heritage
Edited by Martin Van Gelderen and Quentin Skinner
Volume 1: Republicanism and Constitutionalism in Early Modern Europe
0-521-80203-2 Hardback $60.00
Volume 2: The Values of Republicanism in Early Modern Europe
0-521-80756-5 Hardback $60.00

Vico: The First New Science
Edited by Leon Pompa
Cambridge Texts in the History of Political Thought
0-521-38090-1 Hardback $65.00
0-521-38726-4 Paperback $23.00

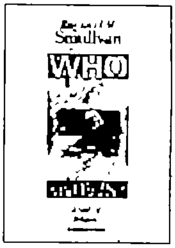
The Cambridge Companion to Duns Scotus
Edited by Thomas Williams
Cambridge Companions to Philosophy
0-521-63205-6 Hardback $65.00
0-521-63563-2 Paperback $23.00

Available in bookstores or from

CAMBRIDGE UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

LATIN AMERICAN PHILOSOPHY

Currents, Issues, Debates
Edited by Eduardo Mendieta
2003
cloth $49.95 paper $22.95



PERFECTIONISM AND CONTEMPORARY FEMINIST VALUES

Kimberly A. Yuracko
2003
cloth $44.95 paper $19.95

WHO KNOWS?

A Study of Religious Consciousness
Raymond M. Smullyan
A lively, stimulating book about religion and mysticism
2003
cloth $39.95 paper $14.95

INDIANA UNIVERSITY PRESS

at bookstores ○ Orders [illegible] ○ iupress.indiana.edu

SURROGATES

Paul Weiss
Foreword by
Robert Cummings Neville
2002 cloth $34.95

Studies in Continental Thought
John Sallis, Editor

ON TRANSLATION

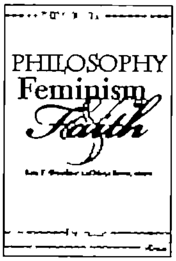
John Sallis
2002
cloth $42.95 paper $19.95

BROKEN HEGEMONIES

Reiner Schürmann
Translated by Reginald Lilly
2002
cloth $89.95 paper $49.95

THE OTHER HUSSERL

The Horizons of Transcendental Phenomenology
Donn Welton
2002 paper $52.95



Indiana Series in Philosophy of Religion
Merold Westphal, Editor

READING HUME'S *DIALOGUES*

A Veneration for True Religion
William Lad Sessions
2002
cloth $49.95 paper $24.95

PHILOSOPHY, FEMINISM, AND FAITH

Edited by Ruth E. Groenhout and Marya Bower
2003
cloth $49.95 paper $24.95

MOTHER / NATURE

Popular Culture and Environmental Ethics
Catherine M. Roach
2003
cloth $55.00 paper $19.95

Prometheus Books

DEMONS OF THE MODERN WORLD
Malcolm McGrath
Foreword by Robert A. Baker, Ph.D.

"*. . . probably the first thorough review of modern demonology . . . superb. Recommended. . . .*"
—*LIBRARY JOURNAL*

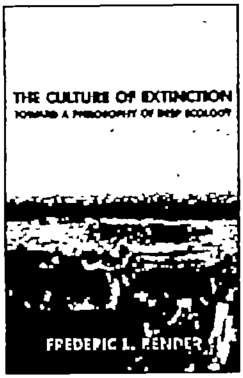
323 pp • ISBN 1-57392-935-2 • HC $32

PHILOSOPHICAL DICTIONARY
Enlarged Edition
Mario Bunge

This indispensable dictionary, directed to general and university audiences, elucidates and evaluates many contemporary philosophical ideas from a humanist and scientifically oriented perspective.

330 pp • ISBN 1-59102-037-9 • PB $24

Humanity Books
an imprint of Prometheus Books



THE CULTURE OF EXTINCTION
Toward a Philosophy of Deep Ecology
Frederic L. Bender

In this probing analysis of the basic cultural assumptions and historical factors that have led to the current dire ecolgical situation, Bender argues that our only chance of avoiding imminent disaster is to recognize and then change the destructive trends in contemporary culture that will surely be our undoing if left unchecked.

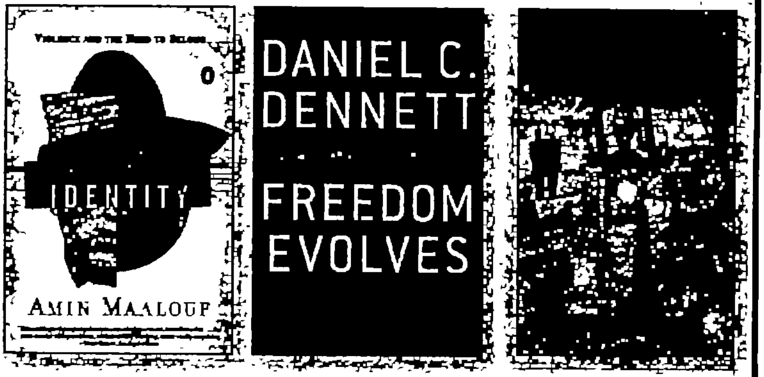
455 pp • ISBN 1-59102-055-7 • HC $65

THE HEIDEGGER-JASPERS CORRESPONDENCE
(1920–1963)
Edited by Walter Biemel and Hans Saner
Translated by Gary E. Aylesworth

This first English translation of the correspondence between two giants of 20th-century German philosophy will be of great interest to philosophers, historians, and anyone intrigued by the Heidegger controversy.

280 pp • ISBN 1-59102-060-3 • HC $59

For exam copies contact editorial@prometheusbooks.com
Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com



FREEDOM EVOLVES
DANIEL C. DENNETT

"Dennett's new book combines, once again, original philosophical thinking, marvelously vivid prose, and extraordinarily lucid argumentation. *Freedom Evolves* does what I would have thought impossible: it says something new about free will and determinism." —Richard Rorty

Viking 352 pp 0-670-03186-0 $25.95

IN THE NAME OF IDENTITY
Violence and the Need to Belong
AMIN MAALOUF

"An eloquent introductory exploration of why, in this age of globalization, we need to abandon our historical idea of identity as a single religious or national alliance." —*The New Yorker*.

Penguin 176 pp 0-14-200257-7 $14.00

SOMETHING FOR NOTHING
Luck in America
JACKSON LEARS

"Nobody is thinking with more spiritedness and subtlety about the roots (and ethical tangle) of American culture and the distinctive American pursuit of happiness."—Susan Sontag

Viking 384 pp 0-670-03173-9 $27.95

FAITH, MADNESS, AND SPONTANEOUS HUMAN COMBUSTION
What Immunology Can Teach Us About Self-Perception
GERALD N. CALLAHAN, Ph.D.

"A fascinating and personal portrait of the immune system. In Callahan's picture, it's the immune system, not our brains, that defines our humanity."—Joe Palca, NPR Sciences correspondent.

Berkley 256 pp 0-425-18852-3 $14.00

LETTER TO A PRIEST
SIMONE WEIL

An intimate letter by one of the twentieth century's greatest spiritual thinkers delves into her struggles with faith and the church.

Penguin 96 pp 0-14-200267-4 $12.00

HAMLET: Poem Unlimited
HAROLD BLOOM

"The indispensable critic on the indispensable writer."—*The New York Review of Books*. In *Hamlet: Poem Unlimited*, our most celebrated critic seeks to uncover the mystery of both Prince Hamlet and the play itself. Examining the major soliloquies, scenes, characters, and action of the play, Bloom limns the enigma at the heart of the drama.

Riverhead 192 pp 1-57322-233-X $19.95

PENGUIN GROUP USA
Academic Marketing Department, 375 Hudson Street New York, NY 10014

THE JOURNAL OF PHILOSOPHY

Calcutta University Library
ALIPORE CAMPUS

VOLUME C, NUMBER 5
MAY 2003

page
221 *A Desire of One's Own* Michael E. Bratman
243 *A Different Conception of Scientific Realism: The Case for the Missing Explananda* James Derden

BOOK REVIEWS
268 Timothy Williamson: *Knowledge and Its Limits* Graham Priest

272 IN MEMORIAM: JAMES J. WALSH
273 NEW BOOKS: TRANSLATIONS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, if the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2003 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 5, MAY 2003

13/8/07

A DESIRE OF ONE'S OWN*

In his 1971 article, "Freedom of the Will and the Concept of a Person," Harry G. Frankfurt[1] pointed to a deep problem in the philosophy of action and sketched an attractive proposal for a solution. The problem arises when we take seriously the idea that you can sometimes have and be moved by desires which you in some sense disown. When such a desire moves you to act the action is an issue of a desire that is, in one straightforward sense, yours—after all, whose desire is it if not yours? But in a deeper way, the desire's motivational impact on your action does not constitute your direction or governance of the action. As Frankfurt put it, though you do what you want you do not act of your own free will.[2] The problem is whether we can make sense of these ideas of—as I will say—ownership and rejection of a desire, without appeal to a little person in the head who is looking on at the workings of her desires and giving the nod to some but not to others. And what we have learned from Frankfurt

---

*Thanks to John Fischer, Harry Frankfurt, Joshua Gert, Carl Ginet, Elizabeth Harman, Terence Irwin, Agnieszka Jaworska, Alfred Mele, Elijah Millgram, Geoffrey Sayre-McCord, T. M. Scanlon, Jeffrey Seidman, Michael Smith, Ralph Wedgwood, Susan Wolf, and audiences at MIT, Cornell, and Florida State University Philosophy Colloquia, the Utrecht Conference on Reasons of One's Own, the Jowett Lecture at the University of Oxford, the Bowling Green Conference on Reason and Deliberation, the University of London Workshop on the Philosophy of Action, the Ethics Workshop at the University of Zurich Ethics Center, and the 2001 Urbino Workshop on Moral Philosophy. I am particularly indebted to thoughtful written comments from John Broome, Peter Railton, Aaron Zimmerman, to extremely helpful conversations with Nadeem Hussain, and to extremely helpful conversations and correspondence with Gideon Yaffe. Much of my work on this essay was supported by a fellowship from the John Simon Guggenheim Memorial Foundation

[1] This JOURNAL, LXVIII, 1 (January 14, 1971): 5–20, reprinted in his *The Importance of What We Care About* (New York: Cambridge, 1988), pp 11–25

[2] The "will" for Frankfurt (in this 1971 article) is, roughly, one's effective desire. My own approach rejects this Hobbesian view

0022-362X/03/0005/221–42

is that our response to this problem has significant implications for our overall view of human agency and autonomy.[3]

I. OWNERSHIP, HIERARCHY, AND THE PLATONIC CHALLENGE

Frankfurt's proposed solution to this problem, sketched in his 1971 article, has come to be called the *hierarchical model*. A desire to *A* (where *A* is something you might do) is yours, in the strong sense we are after, when (roughly) it is a desire you want to move you to act (and so be your "will"). Frankfurt calls this latter, higher-order desire—the desire that the desire to *A* effectively motivate—a *second-order volition*.

Even in 1971, however, Frankfurt saw that this proposal, as just stated, would not work on its own. We need some further story about why certain higher-order desires establish where *you* stand—why they have, so to speak, *agential authority*.[4] And we need to do this without involving ourselves in a criticizable regress. This issue was pressed with insight by Gary Watson[5] in his 1975 reply to Frankfurt's essay; and I have discussed it elsewhere.[6] In those articles, I propose a version of a hierarchical model that draws on my views about planning agency.[7] One of the main ideas is that the relevant higher-order attitudes will include higher-order intentions and policies concerning what considerations are to be treated as reasons in motivationally effective deliberation. These intentions and policies involve distinctive commitments concerning associated forms of practical thought and action, and play central roles in the cross-temporal organization of our temporally extended lives. I argue that such a view provides us with resources for a response to these worries about agential authority.[8]

---

[3] As I will sometimes say, human agents are capable of strong forms of agency. In addition to being purposive agents, human agents have capacities for more complex forms of agency. They are reflective about their motivation and about what is of value. They have long-term plans and policies that organize their practical thought and action over time. Their motivations, value judgments, plans, and policies at any one time have normally been shaped in part over time by ongoing exercises of these capacities. My talk of strong forms of agency is intended to allude to capacities such as these. An adequate philosophy of action needs to help us understand these capacities, their inter-relations, and their relations to self-government and autonomy As I see it, the idea of ownership of desire earns its place by helping us in this project

[4] For this terminology, see my "Two Problems about Human Agency," *Proceedings of the Aristotelian Society*, CI (2001)· 309–26

[5] "Free Agency," this JOURNAL, LXXII, 8 (April 24, 1975): 205–20.

[6] "Identification, Decision, and Treating as a Reason," reprinted in my *Faces of Intention* (New York: Cambridge, 1999), pp. 185–206; "Reflection, Planning, and Temporally Extended Agency," *Philosophical Review*, CIX (2000) 35–61; and "Two Problems about Human Agency."

[7] I develop the planning theory of intention primarily in *Intention, Plans, and Practical Reason* (Cambridge: Harvard, 1987; reprint, Stanford. CSLI, 1999).

[8] The argument for this draws on distinctive ways in which such intentions and

In the present essay, however, I largely put to one side these differences between my view and Frankfurt's versions of a hierarchical theory. Though these differences are significant, and I will at times appeal specifically to my own account, my focus here will be on an issue that arises quite generally for a hierarchical theory. It seems that the relevant higher-order attitudes—whether Frankfurtian desires or higher-order intentions and policies—will themselves normally be to some extent grounded in and constrained by reflection on what one takes to be of value. Indeed, it seems that, normally, if an agent's relevant higher-order attitudes are not to some extent shaped by her evaluative reflections and judgments her agency will be flawed.[9] But this suggests a challenge to the hierarchical account of ownership. The challenge is to explain why we should not see such evaluative judgments—rather than broadly Frankfurtian higher-order attitudes—as the fundamental basis of ownership or rejection of desire: it is when the role of desire in practical thought and action is determined by such evaluative judgments that one's action is self-determined, or autonomous, or performed of one's own free will. Call this the *Platonic challenge*.[10]

One response to this challenge might be to sever the supposed link between Frankfurtian higher-order attitudes and evaluative judgment.[11] My own view, however, is that we should retain some such

---

policies organize our temporally extended lives by way of relevant psychological continuities and connections—see "Reflection, Planning, and Temporally Extended Agency." I say more about why such policies about practical reasoning normally involve hierarchy in my "Autonomy and Hierarchy," *Social Philosophy and Policy*, XX (2003): 156–76.

[9] Eleonore Stump emphasizes a need for some such evaluative grounding in her "Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will," this JOURNAL, LXXXV, 8 (August 1988): 395–420. And see Susan Wolf, "The True, the Good, and the Lovable: Frankfurt's Avoidance of Objectivity," in Sarah Buss and Lee Overton, eds., *Contours of Agency Essays on Themes from Harry Frankfurt* (Cambridge: MIT, 2002), pp. 227–44.

[10] This is one of the challenges posed by Watson's Platonic proposal in his "Free Agency." (Watson, however, qualifies this proposal in his discussion of "perverse cases" in his "Free Action and Free Will," *Mind*, XCVI (1987): 145–72.) Jan Bransen points to a closely related challenge in "True to Ourselves," *International Journal of Philosophical Studies*, VI (1998) 67–85, see p. 76 Michael Ridge has pressed this issue in correspondence, as have Agnieszka Jaworska and Govert den Hartogh in conversation.

[11] This seems to be Frankfurt's view. For example, as Jennifer Rosner has emphasized, in his 1971 article, Frankfurt held that "a person may be capricious and irresponsible in forming his second-order volitions and give no serious consideration to what is at stake" (p. 19n). More recently, Frankfurt says that higher-order "support or rejection of first-order desires need not include, or be based upon, any favorable or unfavorable attitude concerning how desirable or worthy of approval those desires may be"—"Reply to Michael E. Bratman," in Buss and Overton, eds., *Contours of Agency: Essays on Themes from Harry Frankfurt*, pp. 86–90, see p. 87. Frankfurt does,

link while blocking the Platonic challenge. I do not say that such an evaluative backing is, in general, necessary for agential authority or desire ownership or rejection. But I do think that a systematic absence of connection between higher-order Frankfurtian attitude and evaluative judgment would be a breakdown in proper functioning. I want to explain how we can grant this point and still block the Platonic challenge.

II. FEATURES OF FRANKFURTIAN HIERARCHY

Suppose Green has treated you badly, and so you are angry at him and have an associated desire to express this anger.[12] On a Frankfurtian theory you might nevertheless reject this desire. On Frankfurt's 1971 view this rejection would consist primarily of a second-order volition that this desire not move you to action. Other versions of a broadly Frankfurtian hierarchical view are available however, so it would be useful to point to features that will be common to such theories.

I think we can say that on a broadly Frankfurtian view your rejection of your angry desire will have at least the following features. First, it will involve a second-order attitude that is about that desire.[13] Second, this second-order attitude will itself be a conative attitude, in the broad, generic sense of a motivating attitude. Third, this second-order conative attitude will concern certain kinds of further functioning, from now on, of the first-order desire. The content of this second-order attitude will be in this sense forward-looking. Fourth, this forward-looking second-order conative attitude will include in its own functioning the guidance, from now on, of the functioning of the first-order desire. In short: the theory will appeal to a higher-order attitude that is conative, forward-looking in its content, and guiding in its function.

There is also a fifth feature that such theories try to capture. The higher-order, forward-looking, and guiding conative attitude is to constitute—at least in part, and given relevant background conditions—a commitment on the part of the agent concerning the role of the target desire in her own agency: the agent is appropriately settled on this. When we try to say what is needed for such agential commitment, we face issues about agential authority noted earlier.

---

however, note that "[n]eedless to say, it is better for us to care about what is truly worth caring about..."—"On Caring," in his *Necessity, Volition, and Love* (New York: Cambridge, 1999), pp. 155–80, see p. 162.

[12] I leave open the issue whether this desire is in some way essential to the anger.

[13] To keep the discussion manageable, I put to one side yet-higher-order attitudes.

For now, though, it will suffice just to note that hierarchical theories at least aim at capturing some such commitment.

This quintet of features has remained constant throughout many discussions of hierarchical models. For example, in 1987 Frankfurt suggested that the higher-order pro attitude should concern, not directly whether the first-order desire motivates, but whether the first-order desire is "to be endorsed as a legitimate candidate for satisfaction."[14] This involved a change (from the earlier appeal to second-order volitions) in the precise content of the higher-order attitude;[15] but that content nevertheless remained forward-looking, and the attitude retained a downstream guidance function. And Frankfurt's talk in this 1987 essay about decisive commitment makes clear the aim of capturing the kind of agential commitment we have noted.[16]

More recently, in articles to which I alluded earlier, I have tried to understand ownership of a first-order desire for *X* primarily by appeal to a higher-order policy in favor of treating that desire as reason-providing in motivationally effective practical reasoning. Such a policy favors the motivational functioning of the first-order desire by way of practical reasoning whose content involves the thought of *X*, or of the desire for *X*, as a justifying consideration.[17] Despite differences from Frankfurt's accounts, appeal to such policies remains an appeal to a higher-order attitude that is conative, forward-looking in content, and guiding in function. Such policies are intentions that are appropriately general; so they involve distinctive commitments and bring with them distinctive norms of consistency, coherence, and stability over time. In these ways the theory aims to shed light on the kind of agential commitment that is a concern of hierarchical theories.

---

[14] "Identification and Wholeheartedness," in *The Importance of What We Care About*, pp. 159–76, on p. 170. I think Frankfurt's later appeals to satisfaction and to volitional necessity also fit into the cited mold, though they aim at a kind of commitment that—in contrast with the 1987 view—need not involve decision—see Frankfurt's "The Faintest Passion," "On the Necessity of Ideals," and "Autonomy, Necessity, and Love," in his *Necessity, Volition, and Love*.

[15] This change allows Frankfurt to say that one can decide in a particular case not to act on a desire for *X* and still not reject that desire: one might simply give it a lower priority than a competitor in the present circumstance. See Frankfurt's remarks about two different kinds of conflict in "Identification and Wholeheartedness," see p. 170. I discuss this point in "Identification, Decision, and Treating as a Reason," pp. 192, 195–98. See also Joseph Raz, "When We Are Ourselves: The Active and the Passive," in his *Engaging Reason. On the Theory of Value and Action* (New York: Oxford, 1999), pp. 5–21, see p. 18.

[16] "Identification and Wholeheartedness," p. 170.

[17] The agent also needs to be, to use Frankfurt's term, satisfied with the policy I propose a way of understanding such satisfaction in "Reflection, Planning, and Temporally Extended Agency," at pp. 48–50 and pp. 59–60. My proposal there allows for satisfaction with such a policy even in the face of certain strongly felt, first-order desires to the contrary. This feature of the proposal might be challenged, but this is a challenge we need not address here. What is important for present purposes is, rather, that such satisfaction need not in general involve the kind of strong evaluative judgment cited by the Platonic challenge; more on this below.

I do not mean to say that the cited quintet exhausts the features of interest in hierarchical theories. Indeed, I will later emphasize yet a further feature. We can nevertheless use this quintet for an initial characterization of the kind of theory of interest. Our concern, then, will be with hierarchical theories of desire ownership and rejection in which basic theoretical work is done by higher-order attitudes that are conative, forward-looking in content, guiding in function, and purportedly constitute, in context, a relevant commitment on the part of the agent concerning the role of the desire in her agency. Let us call such theories, and the higher-order attitudes they highlight, *Frankfurtian*.[18]

Now, I have maintained that such Frankfurtian attitudes will normally be to some extent grounded in and constrained by reflection on what one takes to be of value. Your higher-order rejection of your angry desire is, let us suppose, anchored at least in part in your view that it is good to be a person who turns the other cheek. The Platonic challenge is that once we note this grounding role of evaluative judgment, our account of what constitutes ownership of desire can dispense with essential appeal to such Frankfurtian attitudes. In the example, perhaps we can say that the fact that the desire to express your anger is not fully your own consists in its failure to cohere with your value judgment in favor of being a person who turns the other cheek. That, anyway, is the challenge.

One might respond by emphasizing that issues of ownership also arise concerning value judgments.[19] It seems, though, that the Platonic challenger can grant this point. She claims that ownership of desire consists, at bottom, in value judgment; to this she can add that the relevant value judgments need to be, in an appropriate sense, the agent's own. But she can claim that for a value judgment to be an agent's own is not for it to be supported by a higher-order Frankfurtian attitude. After all, even a Frankfurtian theory will at some point appeal to features of a Frankfurtian attitude that do not involve a yet-higher-order attitude but help give it agential authority. This is where appeals to, for example, satisfaction and/or a cross-temporal organizing role can take center stage.[20] The Platonic challenger can herself draw on similar resources but apply them to value judgment rather than Frankfurtian higher-order attitude.[21]

---

[18] Such theories can allow that there are cases in which desire leads to action in the absence of any higher-order attitude, pro or con. The claim is only that we should appeal to such higher-order attitudes in a model of strong forms of agency, including self-determination and autonomy

[19] This was Frankfurt's response in conversation. See also J. David Velleman, "What Happens When Someone Acts?" in his *The Possibility of Practical Reason* (New York: Oxford, 2000), pp. 123–43, on p. 134.

For this reason, my strategy here will be to allow the Platonic challenger to appeal to such value judgments, but to argue that in important cases these judgments nevertheless leave open basic issues of desire ownership.

III. WEAKNESS OF WILL AND WATSONIAN PERVERSE CASES

An initial move appeals to cases of weakness of will and the like—cases in which value judgments do not bring with them relevant commitments, and relevant commitments go against one's value judgments. Perhaps I think it strictly better to be a person who forgives and turns the other cheek but nevertheless, in a kind of self-indulgence, allow into my life a willingness to express reactive anger. Though this role of my desire to express my anger diverges from my relevant evaluative judgments, it is not a desire I reject or disown. Indeed, in some versions of such a case—ones Watson calls "perverse cases"—I really am fully behind the expression of such reactive anger. As Watson says, "There is no estrangement here."[22] So, value judgment is one thing, ownership another.

I think this is right as far as it goes; but it does not yet go far enough. After all, it seems that these are not cases of a well-functioning agent. A defender of the Platonic challenge might still offer his view as a view about a well-functioning agent: at least for such an agent, desire ownership and rejection are constituted by value judgment rather than Frankfurtian higher-order attitude. I think, however, that even this qualified Platonic challenge does not work.

IV. THREE EXAMPLES AND AN OBJECTION

Suppose that you and I both like to drink alcohol. We both think that there is some value in drinking alcohol when that is what one wants to do. We have both reflected also on the impact of such drinking on our lives, and we both see that there would also be a certain value in systematically abstaining from alcohol. So far, we are alike. But suppose that you go on reflectively to reject your desire to drink alcohol whereas I do not. In what does your rejection consist? The answer offered by a Frankfurtian theory is that it consists in a higher-order, conative attitude, one that is forward-looking and guiding, and one that partly constitutes a commitment not to build that desire and its targeted activity into your life. A natural picture here is that this Frankfurtian attitude is to some extent grounded in your judgment of the value of abstinence, but it goes beyond that judgment. You arrive at this higher-order attitude, but I do not; yet we share the cited judgments of value. The fact that your lingering desire to drink is one you disown does not consist solely in those judgments, but essentially involves a Frankfurtian attitude.

---

[20] See "Two Problems about Human Agency," pp. 318–19.

[21] Though her explanation of the organizing role of relevant value judgments will differ from the analogue offered within a Frankfurtian theory.

[22] "Free Action and Free Will," p. 150.

Granted, once you arrive at the cited Frankfurtian attitude there may be available a further judgment of value—namely, a judgment that it is good to stick with this commitment.[23] There is room here for disagreement about just how important a value this is. But we can grant the Platonic challenger that such a commitment to abstinence can be sufficiently integral to the agent's life for there to be a distinctive value of living in accord with it. Nevertheless, an appeal to such a value judgment to explain your rejection of your desire for alcohol would not support an effort to dispense with the Frankfurtian apparatus in the account of ownership and rejection.[24]

For a second case, return to you and your anger at Green.[25] Suppose that Green has also treated me badly, and that I too am angry and want to express my anger at Green. Suppose that both of us think that anger, resentment, and their expression are in some ways fitting and appropriate reactions to such untoward behavior. We both also see the good in being a person who turns the other cheek. Where we differ is that after reflection you, in part on the basis of the latter judgment, come to reject your angry desire, whereas I do not. We share a complex evaluative judgment of such anger and its expression; but you go on, and I do not, to a higher-order Frankfurtian commitment not to build the expression of such anger into your life. If your angry desire were to linger, it would be one you disown; not so for me. And it seems that what accounts for this difference is not a difference in value judgment but a difference in commitment, a difference of a sort that a Frankfurtian theory is well positioned to capture.

---

[23] See Raz's important discussion of related ideas in his *The Morality of Freedom* (New York: Oxford, 1986), chapter 14, especially pp. 385–90.

[24] I discuss a related complexity below in section VII.

[25] My thinking about this case benefitted from discussion of related ideas in a meeting of the Bay Area Forum for Law and Ethics (February 2001) in which Samuel Scheffler presented his "Doing and Allowing" (unpublished manuscript). Scheffler too emphasizes the significance to our agency of our need to respond to diverse values, and notes that this is a central theme in work of Raz.

Again, there is the complexity that given your Frankfurtian rejection of your angry desire there is a further judgment of value that may be available—namely, a judgment that it would be good to stick with this commitment. But, again, this is not a way of undermining the significance of the Frankfurtian hierarchy.

For a third case, consider two people faced with a wartime draft. Each, let us suppose, sees value in resisting the call to arms; but each also sees value in loyally serving in one's country's military. Let us suppose that person 1, at the time of decision and after much soul searching, arrives at a forward-looking and guiding Frankfurtian commitment in favor of resisting military service and rejecting his desire to serve: the desire to serve is not to play its normal, end-setting roles in his deliberation and action. Person 2, in contrast, decides in favor of military service. The nagging tug person 1 continues to experience in the direction of loyal military service is one he disowns even though he acknowledges the value in such service. And the fact that he rejects this desire seems to consist in part in his Frankfurtian commitment; for that is what seems to distinguish him from person 2. Granted, once this conscientious objector has arrived at this Frankfurtian commitment, a further judgment of value may be available, one that appeals to the value of being true to this prior commitment. But this does not give us reason to dispense with appeal to Frankfurtian attitudes.

In each of these examples there are relevant judgments of value on both sides of a practical issue; and a Frankfurtian commitment is to some extent grounded in some of those judgments. But the presence of these evaluative grounds does not entail that there is no further work to be done by the Frankfurtian commitment. In some such cases it is only when one arrives at a Frankfurtian commitment that one has, in the relevant sense, taken a stand with respect to the issues raised for one's life.

This suggests an answer to the Platonic challenge. We can accept that in a well-functioning agent Frankfurtian attitudes will normally be to some extent grounded in and constrained by judgments of value that do not themselves appeal to those Frankfurtian attitudes. We can, nevertheless, still reject the claim that these value judgments by themselves in general fully constitute the basis of ownership and rejection.

There is, however, an objection to this. Consider our conscientious objector. It might be claimed that if he is rationally to arrive at a Frankfurtian rejection of his desire to serve in the military, he must judge not only that resistance has value but also that resistance is strictly better than military service, and perhaps also that it is strictly better to be a person who resists. And similarly in the other examples. Once such strict comparative value judgments are available, however, we can appeal to them, rather than Frankfurtian higher-order attitudes, in a Platonic account of ownership and rejection. Or so the Platonic challenger might claim.

If this is to support the Platonic challenge, the cited strict comparative value judgments must not themselves make essential appeal to prior Frankfurtian attitudes. It seems to me, however, that even in a well-functioning agent, desire ownership or rejection does not in general require the backing of such independent strict comparative value judgments. I proceed to explain why.

V. THE WILL

A number of years ago Donald Davidson[26] proposed that intending is an "all-out" evaluative judgment. In my 1985 discussion I interpreted this as the claim that to intend to *A* is to judge that *A* is strictly better than its relevant alternatives.[27] I then noted that when we see intention as such a strict comparative evaluation we face a Buridan problem; for there are cases in which one decides in favor of one of several conflicting options even though, prior to one's decision, one only judges that the option one decides on is as good as (but is not strictly better than) its relevant alternative.[28]

An analogous point can be made in the present context. Perhaps, prior to his decision, our conscientious objector thinks that military service and resistance to such service really are, everything relevant considered, each as good as the other. Still, life must go on. One way for him to settle the issue would be to settle on a Frankfurtian commitment that rejects his desire for military service as no longer fully his own.[29]

Granted, he might simply make a decision about what to do on the present occasion, and not arrive at a higher-order commitment that rejects his desire to serve.[30] Or he might only decide to give military service a lower priority in his practical reasoning, but not reject his desire for military service as no longer his own. Given the significance to his life of the decision he faces, however, one intelligible response might be not simply to decide not to enlist, and not simply to give one option a lower priority, but to decide quite generally not to treat his own participation in loyal military service as a consideration of positive significance in his own life.[31] Still, one may think that such cases of apparent evaluative ties are rare. So let us extend the point. The next place to look will be cases in which the agent is, prior to his decision, uncertain about which course of action, or way of life, is better.[32] This, too, might be the plight of our conscientious objector. Yet here, again, he really must settle the issue and get on with his life. And he might do this by way of arriving at a Frankfurtian commitment that rejects his desire for military service.

---

[26] "Intending," originally published in 1978 and reprinted in his *Essays on Actions and Events*, second edition (New York: Oxford, 2001), pp. 83–102.

[27] See my "Davidson's Theory of Intention," reprinted in *Faces of Intention*, pp. 209–24. I argued there that if we were instead to cite a weaker evaluative judgment—say, that there is something good about *A*, or that *A* is as good as its alternatives—we would run into difficulties concerning the rational agglomerativity of intention

[28] See "Davidson's Theory of Intention," and Edna Ullmann-Margalit and Sidney Morgenbesser, "Picking and Choosing," *Social Research*, XLIV (1977): 757–85.

[29] Rejecting the desire need not involve a commitment to getting rid of it; the commitment is, rather, to rejecting relevant functioning in practical reasoning and action. As Elizabeth Harman emphasized in conversation, it might even be important to the agent to retain the desire while blocking such functioning.

Finally, an agent might think that there is something of value in both alternatives but that there is no further, relevant comparative evaluation to be made—at least until he makes a relevant commitment. This, too, may be the plight of our conscientious objector. He may think there is value in resistance, and value in loyal military service, but there is, at least prior to his relevant commitment, no further truth about which is better. Still, he must get on with his life.

What these cases have in common is that the agent's value judgments by themselves underdetermine his stance in response to the practical issues raised about how he is to live.[33] For a well-functioning agent in this predicament Frankfurtian commitments can—and, indeed, need to—go beyond prior value judgment. Such commitments do not require the backing of strict comparative value judgments that do not themselves make essential appeal to such commitments.

---

[30] As Aaron Zimmerman has emphasized.

[31] As I would put it, the relevant decision would be a decision not to treat this consideration as a reason in his relevant practical reasoning.

[32] Thanks to Nadeem Hussain for emphasizing this case.

[33] Consider Stuart Hampshire's description of a person who makes a decision in the face of a certain kind of conflict: "he will probably recognize that his choice of a way of life is undetermined by the arguments that support his decision. . He often would not wish to say that his decision on the right course is objectively right, in the sense that in making the decision he is at the same time claiming universal agreement for it.... But the word 'commitment' carries a more positive implication, the implication is that he has himself recognized that his choice is undetermined by the reasons that support it, and, secondly, that he accepts responsibility for the choice as being his and his alone, without the support of any external authority"—*Two Theories of Morality* (New York: Oxford, 1977), p. 52. David Wiggins points to related ideas in his discussion of the "doctrine of cognitive underdetermination"—see his "Truth, Invention, and the Meaning of Life," in *Need, Values, Truth: Essays in the Philosophy of Value*, third edition (New York: Oxford, 1998), pp. 87–137, especially p. 124. See also Wolf, "The True, the Good, and the Lovable: Frankfurt's Avoidance of Objectivity," at pp. 234–35, Richard Holton "Intention and Weakness of Will," this JOURNAL, XCVI, 5 (May 1999): 241–62, at p. 245; and Robert Nozick's suggestive discussion in *Philosophical Explanations* (Cambridge: Harvard, 1981), pp. 446–50. (I discuss these views of

A by-now-familiar complexity is that, once the agent arrives at a Frankfurtian commitment, this may bring to bear the value of living in accord with such commitments. But the Platonic challenge cannot appeal to that further value judgment—a judgment that itself appeals to a prior, distinct Frankfurtian commitment—to support the claim that such commitments are no essential part of desire ownership.

Could such Frankfurtian commitments, without the backing of strict comparative value judgments that do not themselves make essential appeal to such commitments, nevertheless have agential authority? I think the answer is yes. This is because we can expect such commitments normally to be stable, and reasonably so, and to play appropriate roles in the agent's practical thought and action over time. On my version of such a theory, for example, these commitments consist primarily in higher-order policies concerning the role of various desires in one's motivationally effective practical reasoning. Such policies will normally have a distinctive stability and will support the cross-temporal organization of the agent's practical thought and action.[34] (I return to this matter below in section VIII.)

In the kind of case of underdetermination I am emphasizing, one may arrive at a Frankfurtian commitment by way of reflection that is responsive to a wide range of relevant thoughts, feelings, and inclinations. And there is probably an important role here for what John Dewey called "dramatic rehearsal (in imagination) of various competing possible lines of action."[35] The claim is not that one's arriving at such a commitment need be simply a matter of unreflective, brute picking.[36] In reflectively arriving at the commitment, however, one may well go beyond one's value judgments that are independent of such commitments.

A further complexity is that the agent may arrive at a new Frankfurtian commitment in a way that draws on connections or parallels or analogies with other, pre-existing Frankfurtian commitments. This

---

Nozick in "Nozick on Free Will," in David Schmidtz, ed., *Robert Nozick* (New York: Cambridge, 2001), pp. 155–74.) Keith Lehrer also highlights the relation between hierarchical theories and a kind of underdetermination in his *Self-Trust: A Study of Reason, Knowledge, and Autonomy* (New York: Oxford, 1997), chapter 4.

[34] See my "Reflection, Planning, and Temporally Extended Agency."

[35] Dewey, *Human Nature and Conduct: An Introduction to Social Psychology* (New York: Random House, 1957), p. 179.

[36] See Ullmann-Margalit and Morgenbesser, "Picking and Choosing."

may sometimes have a structure analogous to an appeal to precedent in certain forms of legal reasoning.[37] And perhaps this is one way of understanding Charles Taylor's appeal, in such contexts, to "our sense of the shape of our lives."[38] But such reasoning does not support a Platonic rejection of the role of those prior commitments.

Yet another complexity concerns the significance to this story of an act of decision. It has been useful to think of our examples as involving a moment of decision; but this is not essential to the point I want to make. What is central is the idea of a kind of attitude that, in this context, constitutes the relevant form of agential commitment. Decision is sometimes a way of arriving at such an attitude; but such an attitude may also emerge from a more complex, inchoate and extended process of reflection. Further, if one were to think one has decided to reject a certain desire, but one did not thereby actually arrive at a relevant attitude, then one would not have arrived at the relevant commitment.[39]

VI. COMPARATIVE VALUE JUDGMENT AND INTERSUBJECTIVITY

What conception of value judgment makes the most sense of these observations? This is a big issue. Here I limit myself to remarks about comparative value judgments and intersubjectivity, remarks that are intended to get at a feature of such judgments that will be acknowledged by a wide range of theories.

It is natural to see a judgment that *X* is better than *Y* as not merely expressing a personal ranking, but as being in the domain of the intersubjective. There are different ways of trying to articulate this idea. We might say that a judgment that *X* is better than *Y* is made from, as Hume would say, a "common point of view,"[40] or is at least in part a judgment that from a common point of view one would favor *X* over *Y*. Or we might say that such a judgment tracks a certain kind of interpersonal convergence. Or we might see such a judgment as, in part, the expression of an associated demand on others. And other views are possible. Here, I want to remain neutral concerning how best to articulate this connection between value judgment and intersubjectivity, for I think that the main point I want to make will be available on any acceptable interpretation of this connection.[41]

---

[37] See Nozick, *Philosophical Explanations*, p. 297.

[38] Taylor, "Leading a Life," in Ruth Chang, ed., *Incommensurability, Incomparability, and Practical Reason* (Cambridge: Harvard, 1997), pp. 170–83, at p. 183.

[39] Cf. Frankfurt, "The Faintest Passion," p 101.

[40] *An Inquiry Concerning the Principles of Morals*, section IX.

[41] In particular, I want to remain neutral in the debate between cognitivist and expressivist treatments of this connection.

It will help keep things manageable, however, to work with a particular interpretation. So let us consider an interpretation of this idea suggested by work of Michael Smith. On Smith's view, a judgment that *X* is desirable is a judgment about "convergence in the desires of fully rational agents" concerning "what is to be done in the various circumstances in which they might find themselves."[42] A corresponding view about comparative evaluations would be that a judgment that *X* is better than *Y* in certain circumstances involves a judgment that such fully rational advisors would converge on an all-considered preference for *X* over *Y* in those circumstances. On such a view, for our conscientious objector to judge that in his circumstances resistance is better than military service, he needs to judge that fully rational advisors would converge on an all-considered preference that people, in such circumstances, resist rather than join the military.

Will our conscientious objector think this? The answer will likely depend on how the cited circumstances are described. Our conscientious objector may think as follows: "If the cited circumstances are as my circumstances were before I decided in favor of resistance, then there seems little reason to expect such convergence among all rational advisors specifically in favor of resistance.[43] If, however, the circumstances include the agent's Frankfurtian commitment to resistance, then perhaps I would expect that rational advisors would converge on an all-considered preference in favor of the agent's resistance."[44]

This may be over-simplified. Perhaps there would be reason to expect such convergence concerning circumstances that do not include a commitment specifically to resistance, so long as those circumstances include certain other, related Frankfurtian commitments of the agent. Even if our conscientious objector were to recognize this complexity, however, he still might have no relevant comparative evaluative judgment that satisfies the intersubjectivity constraint, does not appeal to prior Frankfurtian commitments, and yet settles the practical issue.

This suggests a general point. Judgments of comparative value are subject to an appropriate intersubjectivity constraint. Here we have considered an interpretation of this constraint that sees the judgment as concerned with a convergence in the preferences of fully rational advisors; and one might question this particular story. But it seems in any case that some version of the connection with intersubjectivity will apply to the comparative evaluative judgments relevant here. And what our case of the conscientious objector suggests is that in some cases an agent's comparative evaluative judgments that respect relevant intersubjectivity constraints, but do not appeal to prior Frankfurtian commitments, will be by themselves too weak to settle the practical issue that is at stake. In such cases, Frankfurtian attitudes can settle the practical issue; but—and this is the further feature of such attitudes to which I alluded earlier—they do not seem to be subject in the same way to corresponding intersubjectivity constraints. After all, the primary role of Frankfurtian commitments is to settle where *that agent* stands. This will, I think, require something roughly along the lines of what Frankfurt calls "wholeheartedness" on the part of the agent.[45] It does not, however, seem to require an assumption by that agent that all rational advisors would converge on (or an injunction by that agent that everyone is to converge on) the very same stance.[46] This divergence in relevant intersubjectivity constraints on comparative value judgment, on the one hand, and on Frankfurtian commitment, on the other, is at the heart of my response to the Platonic challenge.[47]

---

[42] *The Moral Problem* (Cambridge: Blackwell, 1994), p. 173. See also Michael Smith, "In Defense of *The Moral Problem*: A Reply to Brink, Copp, and Sayre-McCord," *Ethics*, cviii (1997): 84–119, at pp. 88–90.

[43] Cf. Wiggins: "Often we have to make a practical choice that another rational agent might understand through and through, not fault or even disagree with, but (as Winch has stressed) make differently himself..."—"Truth, Invention, and the Meaning of Life," p. 126.

[44] Though see below, section vii.

Return to the case in which our conscientious objector judges that there is value both in resistance and in loyal military service, but no further truth about which is better. He might reach this judgment concerning circumstances that do not yet include relevant Frankfurtian commitments one way or the other. After all, those may well be the circumstances he is in prior to his decision. He may think that if such commitments are not part of the circumstance there is no reason to expect appropriate convergence, or its analogue, specifically on one of the options. And he may then proceed, on reflection, to arrive at a Frankfurtian commitment that goes beyond his current evaluative judgments.

---

[45] See especially Frankfurt's "The Faintest Passion." I do not try here to settle the further issue of precisely how to understand such wholeheartedness. For present purposes, the important point is only that wholeheartedness does not require the cited kind of intersubjectively accountable strict comparative evaluation.

[46] So our conscientious objector may reflectively decide to reject his inclination towards loyal military service and yet not think, as J. David Velleman puts it, that this "is what anyone would think in response to the relevant practical question, and would think that anyone would think, and so on"—see Velleman's "The Voice of Conscience," *Proceedings of the Aristotelian Society*, XCIX (1998): 57–76, at p. 69. And compare Frankfurt's observation that "it is not very likely...that what each of us considers most important to himself is exactly the same"—"Autonomy, Necessity, and Love," at p. 132

[47] And it is related to Allan Gibbard's distinction between an "existential commitment" and acceptance of "a norm as a requirement of rationality"—see *Wise Choices, Apt Feelings* (Cambridge: Harvard, 1990), pp 166–70. See also (in an essay to which Gibbard here refers) P. F. Strawson's discussion of "the natural diversity of human ideals"—"Social Morality and Individual Ideal," in his *Freedom and Resentment and Other Essays* (New York: Methuen, 1974), pp. 26–44, at pp. 42–43.

VII. REFLEXIVITY AND THE VALUE OF LIVING IN ACCORD WITH ONE'S COMMITMENTS

We now need to consider a further Platonic proposal, one that draws on the value of living in accord with one's commitments.[48] The proposal is that ownership of a desire for *X* consists in a *reflexive* value judgment roughly along the lines of:

> (P) It would be strictly best, in part because of this very judgment that (P), to be a person who acts on the desire for *X* in relevant circumstances.

One can arrive at such a reflexive judgment even though one does not think that, prior to and independently of that judgment, it would be best to be a person who acts on the desire for *X*. The reflexive judgment, so to speak, sees itself as part of what determines what would be best. And the proposal is that the agent's relevant commitment consists in such a reflexive value judgment.

How can this evaluative judgment itself help determine what would be best? The idea is that the existence of the judgment brings to bear the value of acting in accord with such judgments. But why does that have value? Because the judgment is the agent's commitment, and there is value in living in accord with one's commitments.

This proposal remains vulnerable to concerns about weakness of will and the like; for even if one arrives at (P) one may still fail, out of weak will or "perversity," to endorse the cited desire. However, the Platonic challenger might still propose that, barring such breakdowns, desire ownership is in general constituted by judgments along the lines of (P).

This proposal does provide a more subtle treatment of some examples. Nevertheless, it remains in tension with our observations about intersubjectivity. Granted, intersubjectivity constraints may be easier for a judgment that (P) to satisfy given that the relevant circumstances now include the making of the judgment itself. It seems, however, that an agent might own or identify with a desire without making such an intersubjectively accountable value judgment, and without being guilty of weakness of will or the like. We can see the point by returning to our Smith-like interpretation of the relevant intersubjectivity: the agent might identify with her desire without supposing that there would be relevant convergence among rational advisors even with respect to the situation in which she identifies with that desire. She may have little confidence that there would be such convergence; she may even believe, to the contrary, that reasonable advisors would indeed diverge in their preferences concerning whether or not, given that one has identified with this desire, one is to act on it.

---

[48] This proposal is broadly in the spirit of a suggestion of Velleman about how Davidson's theory of intending might be developed to respond to Buridan cases. Velleman made this suggestion in his contribution to a symposium on my *Intention, Plans, and Practical Reason* at the 1988 meetings of the Central APA. A related idea is in Wolf, "Two Levels of Pluralism," *Ethics*, CII (1992), 785–98, at p. 796.

Suppose, for example, the desire in question is the agent's desire not to help people in need. Suppose she identifies with this desire after reading a novel by Ayn Rand. She might nevertheless recognize that this is quite controversial and so not be confident that all rational advisors would converge. She might even think that some (though not all) such advisors would favor her acting rather on a desire to help, even given that she identifies with her desire not to.[40] This need not, however, prevent *her* from identifying with her desire not to help. And this can be explained by a Frankfurtian theory: for neither the absence of a belief in convergence, nor even a belief in nonconvergence, need stand in the way of a Frankfurtian attitude that endorses the desire not to help. Either of these would, however, stand in the way of her arriving at a relevant version of (P)—at least, they would stand in the way if she were reflective about the Smith-like convergence constraints we are now supposing to apply to such a value judgment. So, our Randian agent may endorse her desire not to help even though she does not arrive at a version of (P) concerning that desire. She aims at fashioning a coherent life of her own, within constraints of the good. So long as she does not think it strictly better to abandon her commitment to selfishness, she can shape her life of selfishness within these constraints (as she understands them).

A response might be that in such cases the value of living in accord with one's commitments is always sufficient decisively to tilt the evaluative scales in favor of living in accord with one's commitments. So our agent is at least in a position to arrive at a version of (P) concerning her desire not to help, since she is in a position to see that her arriving

[40] Such an advisor would favor something analogous to what Nomy Arpaly and Timothy Schroeder call "inverse akrasia"—see their "Praise, Blame and the Whole Self," *Philosophical Studies*, XCIII (1999): 161–88.

at this judgment ensures its truth. This does not, however, seem in general quite right. That one's commitments organize and structure one's life over time is, I think, part of what gives them authority to speak for the agent. It does not follow, however, that living in accord with those commitments either has, or must be thought of by the agent as having a value that, in such cases, is always sufficient decisively to tilt the intersubjective evaluative scales. Authenticity has its limits, even in the eyes of an authentic agent.[50]

VIII. A FINAL PLATONIC CHALLENGE

There is, then, reason to see Frankfurtian attitudes as essential to important cases of desire ownership in which a relevant strict comparative value judgment is not available to a well-functioning agent. Further, once these Frankfurtian structures are in place we have the resources for a plausible model of the kinds of weakness of will and Watsonian "perversity" noted earlier. For me to build into my life a commitment to the expression of reactive anger—even though I think it best not to—is for me to arrive at a Frankfurtian commitment to my angry desires despite my value judgment to the contrary. My ownership of those desires is constituted by my Frankfurtian commitment, not my evaluative judgment. Yet further, we can extend the account to agents who, in certain contexts, simply refrain from making intersubjectively accountable judgments about the good, but nevertheless put themselves fully behind certain desires by way of relevant Frankfurtian attitudes.[51]

So we have a fairly uniform story to tell about ownership of desire in underdetermination cases, in cases of weak will and "perversity," and in cases in which the agent refrains from relevant judgments about the good. Further, we can see the risk of Watsonian perversity as built into the very psychic machinery that enables us to respond, as we frequently must, in underdetermination cases.[52] But now we need to ask: What about those cases in which, by the agent's own lights, the evaluative judgment does fully settle the issue of ownership of relevant desires, and on the basis of this the agent is fully behind, or fully rejects, those desires? Perhaps it is clear to you that it is a very good thing to build mercy into your life, and a very bad thing not to; and that is why you are fully behind your desires to act mercifully. Again, perhaps—in contrast with our earlier case—you see no value at all in reactive anger and its expression, and that is why you fully reject your angry desires. In each case, as you see it, your ownership or rejection is not a response to underdetermination by value, but rather to an important, determining value. Call such cases *value-judgment-determined* cases. Does the possibility of value-judgment-determined cases challenge the centrality of Frankfurtian attitudes to desire ownership and rejection?

---

[50] So the agent in our example may believe, roughly, that prior to her identification neither selfishness nor selflessness was superior to the other, that her selfishness after her identification is superior to her selfishness prior to her identification, but that even after her identification neither her selfishness nor her selflessness would be superior to the other. Raz would say that by her lights her case bears "the mark of incommensurability"—see *The Morality of Freedom*, pp. 325–26. Ruth Chang would offer a different way of conceptualizing such a case—see her "Introduction" to *Incommensurability, Incomparability, and Practical Reason*, pp. 23–27. What matters for my purposes here, however, is only that our agent does not arrive at a relevant instance of (P) even though she identifies with the selfish desire.

[51] Here I have benefitted from discussion with Sayre-McCord. See also Frankfurt, "Reply to Susan Wolf," in *Contours of Agency: Essays on Themes from Harry Frankfurt*, pp. 245–52, at p 249.

Well, it is common ground at this point in the discussion that Frankfurtian commitments are involved in a wide range of cases. With these Frankfurtian structures in place it will be plausible to infer that even in value-judgment-determined cases the ownership or rejection of desire involves a Frankfurtian commitment, albeit a Frankfurtian commitment that, by the agent's lights, is made uniquely reasonable by relevant value judgment.[53] By focusing our philosophical microscopes on cases of underdetermination—as well as on cases of weak will, "perversity," and simple absence of evaluative judgment—we uncover Frankfurtian structures that we then infer to be generally involved in desire ownership or rejection. In this way we arrive at a fairly uniform model of desire ownership and rejection across all cases.[54]

---

[52] Here I have benefitted from related remarks of Velleman in his "What Good Is a Will?" (unpublished manuscript). See also R. Jay Wallace's remarks about "a hazardous by-product of the capacity for self-determination" in his "Normativity, Commitment, and Instrumental Reason," *Philosophers' Imprint*, I, 3 (2001): 1–26, at p. 10.

[53] Indeed, it is important to note that even when the agent does judge it strictly best to build, say, mercy into his life there may well still be significant underdetermination of just how important this is to be in his life. So there will continue to be a need to respond to this underdetermination.

[54] As I say, this inference is plausible; but perhaps it is not required. An alternative picture might be that for certain kinds of agents all there need be in a value-judgment-determined case is the value judgment and its guidance of the role of relevant desire in action. For such agents in such cases the value judgment on its own constitutes the relevant commitment at the heart of ownership or rejection. (Both Frances Kamm and Derek Parfit pointed toward such a view in conversation.) But even this alternative picture would grant the significance of Frankfurtian attitudes for all agents in underdetermination cases, as well as for many other cases of desire ownership for some kinds of agents. The crucial issue for present purposes will then be whether these Frankfurtian cases are somehow less full-blown cases of ownership or rejection. And that is the issue I proceed to address in the main text.

But at this point the Platonic challenger might still propose that it is only in value-judgment-determined cases that there is, as it were, full-blown desire ownership or rejection. The other cases I have emphasized are possible, but as cases of desire ownership or rejection they are, so to speak, second (or, in breakdown cases, third) best. So in the end—and this is the final Platonic challenge—we cannot avoid a Platonic picture of the centrality of evaluative judgment, rather than Frankfurtian attitude, to full-blown desire ownership or rejection.

Why believe this? The underlying idea is that value-judgment-determined cases have a feature that is not present in the other cases I have emphasized, a feature that is fundamental to full-blown desire ownership or rejection. What feature? The natural candidate here is a kind of reasonable stability. Desire ownership or rejection involves the agent's stable commitment concerning the desire. And a Platonic challenger might claim that ownership or rejection that is value-judgment-determined will be rationally stable, whereas ownership or rejection anchored in commitment in the face of underdetermination will not.

I think, however, that this last claim rests on a mistake; for commitments in the face of underdetermination by value judgment may well be stable, and reasonably so. Your commitment to resist military service may have arisen in a context of underdetermination by value judgment. Once your commitment is there, however, you may well be, as Frankfurt would say, wholehearted about it, and in such a case you will normally need some further reason to change it.[55] Normally, you will not have such a further reason. After all, the evaluative case on the other side was, by your lights, already less than conclusive; and once your commitment is in place, values of integrity and the like can now argue in its favor.[56] Your commitment can be reasonably settled and stable, and play its characteristic roles in your practical thought and action, even in the face of underdetermination by intersubjectively accountable value judgment. So your commitment can have, as I have put it, agential authority. So Frankfurtian ownership or rejection of desire in such underdetermination cases need not be a second-class citizen. And that answers the final Platonic challenge.[57]

---

[55] On my account this is a special case of the general phenomenon that prior intention-like commitments can reasonably function as a stable (though defeasible) default, given their central role in cross-temporal organization. See, for example, my "Shared Valuing and Frameworks for Practical Reasoning," in Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith, eds., *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (New York: Oxford, forthcoming). (Shelly Kagan, in correspondence, suggested a parallel here with the judicial doctrine of *stare decisis*.) See also Barbara Herman's insightful characterization of Frankfurt's views in her "Bootstrapping," in *Contours of Agency: Essays on Themes from Harry Frankfurt*, pp. 253–74, at p. 259. In "Are Intentions Reasons? And How Should We Cope with Incommensurable Values?"—in Christopher W. Morris and Arthur Ripstein, eds., *Practical Rationality and Preference: Essays for David Gauthier* (New York: Cambridge, 2001), especially pp. 114–19—John Broome explores the potential role of prior intentions in incommensurability cases; but Broome is skeptical about the idea that intentions and the like have the cited kind of reasonable stability.

[56] Though see section VII for relevant qualification.

IX. CONCLUSION

I think, then, that the Platonic challenge does not succeed. In at least a wide range of cases, ownership or rejection of desire goes beyond intersubjectively accountable value judgment and involves distinctive Frankfurtian commitments concerning whether and how to build that desire, and what it is for, into one's practical life.[58]

Now, in the cases of rejection of desire to which Frankfurt originally drew our attention—his unwilling addict, for example—it seems plausible to say that the rejection goes with a judgment that what the desire is for has no positive value, and perhaps also with distinctive feelings of alienation and the like. I have, in contrast, not focused on such feelings and have emphasized cases in which an agent rejects a desire even though she takes it that what the desire is for has some value. Perhaps she sees the value of the pleasures of alcohol, or—to change examples—of sexual activity, but is herself committed to a life of abstinence in which these pleasures are given no weight. On my account, she thereby rejects her desires for these pleasures even though she sees these pleasures as having a certain value. This does not mean that she is completely insensitive to the value of alcohol, or of sex. She may, for example, still buy her friend a drink of wine (but not a dose of heroin). She may respect her friend's way of life, while being clear that this way of life is not her own. In buying her friend a drink she is guided by her commitment to building friendship, and a certain kind of respect for and toleration of her friend, into her own life. But this is not to build into her life her own pursuit of the pleasures of alcohol.[59]

---

[57] My remarks in this paragraph have benefitted from discussions with Nadeem Hussain.

[58] A further issue is whether a full theory will need somehow to limit the etiologies of these Frankfurtian commitments. Perhaps we should say that at least in central cases of ownership the commitments are to some appropriate extent an issue of the on-going exercise of the kinds of basic capacities cited in footnote 3. We do not, however, need to settle this here.

[59] My remarks in the second half of this paragraph owe a great deal to probing queries from Terence Irwin, and to help in responding to those queries provided by Carl Ginet.

In classifying these different cases—including both the unwilling addict and the person committed to abstinence—as ones that involve rejection I am, then, extending this idea beyond the cases to which Frankfurt's early discussions most directly call our attention. What all these cases have in common, however, is that the agent is committed not to give the target desire, and what it is for, certain central roles in her practical life. If she nevertheless, in relevant ways, reasons from and/or acts on that desire, she thereby violates her commitment concerning (to return to Taylor's phrase) the shape of her life. In extreme cases this would amount to a kind of self-betrayal, or failure of self-respect, or self-esteem.[60]

This commonality seems to me to justify a notion of rejection that applies to all such cases. One might perhaps balk here and claim that, despite this commonality, only the first kind of case—in which the agent sees no value at all in what she desires, and perhaps also experiences feelings of alienation—merits the label of rejection. But so long as one grants the commonality I have emphasized the dispute may be merely verbal. The important point is that appeal to this commonality helps clarify structures involved in strong forms of human agency and autonomy; and it does this in ways that are responsive to important phenomena of underdetermination by value judgment. In my judgment, it is such structures toward which our concerns with ownership and rejection of desire are most fruitfully directed.

MICHAEL E. BRATMAN

Stanford University

---

[60] Concerning self-esteem, see Frankfurt's "Reply to Barbara Herman," in *Contours of Agency: Essays on Themes from Harry Frankfurt*, pp. 275–78, at pp. 277–78.

A DIFFERENT CONCEPTION OF SCIENTIFIC REALISM: THE CASE FOR THE MISSING EXPLANANDA

This paper is concerned with the mathematical laws, or theories, of empirical science. We begin with Bas C. van Fraassen's[1] portrayal of scientific realism: "Science aims to give us, in its theories, a literally true story of what the world is like; and acceptance of a scientific theory involves belief that it is true." Of this portrayal, van Fraassen says that "it is quite minimal, and can be agreed to by anyone who considers himself a scientific realist" (*ibid.*, p. 8). I am one scientific realist who cannot agree with this portrayal. But before we proceed, we must settle an issue that might derail the project before it gets a fair hearing.

I. IDEALIZED MATHEMATICAL LAWS ARE VACUOUSLY TRUE

The realist position developed in this article comes about through the recognition of the role of ideal conditions in scientific theorizing. This is best seen through an example in which I make a distinction between mathematical laws and their equations. Thus, the ideal gas *equation* is: $PV = RT$.[2] But the ideal gas *law* is more than this equation. Many idealizing assumptions known to be false are used in constructing or deriving the ideal gas law (and hence its equation). Some of the assumptions for the molecules of a gas in a given container are:

(G_1) The molecules are perfectly spherical.
(G_2) The molecules are perfectly elastic in their collisions with each other and the sides of the container.
(G_3) There are no attractive forces exerted by the molecules or by the sides of the container.
(G_4) The volumes of the molecules are "negligible."

As a first approximation, the ideal gas *law* is the following:

(I) If G_1 & G_2 & G_3 & G_4, then $PV = RT$.

I say "approximation" because the (almost) final analysis of the ideal gas law is the following:

(II) $(u)(z)(x)(y)(w)\{[[u$ is a gas & z is a vessel & u is contained in z & u is composed of molecules all of which are perfectly elastic & u

---

[1] *The Scientific Image* (New York: Oxford, 1980).

[2] Throughout I am taking the number of moles of the gas to be one.

0022-362X/03/243–67

> is composed of molecules all of which have negligible volumes & u is composed of molecules all of which are perfectly spherical & u is composed of molecules all of which exert no forces on one another or with the sides of the container except during collisions] & [$x =$ the numerical value of the pressure of u & $y =$ the numerical value of the volume of z & $b =$ the numerical value of the universal gas constant & $w =$ the numerical value of the absolute temperature of u]$\} \mapsto (xy = bw)\}$.[3]

One sees from (II) that the idealizing conditions are *constitutive* of the laws and theories for which they are invoked. Such conditions, when expressed in the form of singular statements, are known to be false. For example, the statement that the molecules in this container here and now are perfectly spherical is false. Also, the predicates occurring in the antecedent of the generalized conditional are many of them unsatisfied, for example, 'x is a perfectly spherical molecule'. Because of the material conditional, when at least one of the idealizing predicates is unsatisfied, the generalized conditional of which it is a part is true in first order logic. The mathematical laws of empirical science on this account are thus generalized conditionals that are true and known to be so prior to any testing.[4]

II. WHAT ABOUT COUNTERFACTUALS OR RELEVANCE LOGICS?

So, the potentially derailing issue here is whether it is appropriate to take the mathematical laws, as opposed to the equations, to be (vacuously) true. Some will prefer to pursue a relevance logic here, while others will want to pursue theories for counterfactual conditionals. I have chosen neither of the latter approaches, and instead have decided to take as the project one of determining when to accept and when to reject the vacuously true mathematical laws. I suspect that the task of finding criteria for accepting or rejecting the vacuously true laws is an easier one than that of building a general relevance logic or a general counterfactual theory because of the restriction in scope. I also suspect that any criteria I might offer in my project will very likely have to be respected by either of the other two attempts at theories. For example, in the absence of a counterfactual theory, I am betting that my acceptable vacuous mathematical laws will serve as a basis for intuitively acceptable counterfactuals.

---

[3] With (II) the analysis is still not complete, since we have not introduced the units for pressure, temperature, volume, and the gas constant, and since there are more ideal conditions beyond the four above. See Francis W. Sears, *Thermodynamics, The Kinetic Theory of Gases, and Statistical Mechanics* (Reading, MA: Addison-Wesley, 1956), pp. 201–04, for additional assumptions.

[4] They also contain pure mathematical equations as their consequents. For this type of analysis, see W. F. Barr, "A Syntactic and Semantic Analysis of Idealizations in Science," *Philosophy of Science*, xxxviii (1971): 258–72; Leszek Nowak, "Laws of Science, Theories, Measurement: Comments on Ernest Nagel's *The Structure of Science*," *Philosophy of Science*, xxxix (1972): 533–48, Nowak, *The Structure of Idealization: Towards a Systematic Interpretation of the Marxian Idea of Science* (Boston· Reidel, 1980); Nowak, "The Idealization Approach to Science: A Survey," in Jerzy Brzezinski and Nowak, eds., *Idealization III: Approximation and Truth* (Atlanta: Rodopi, 1992), pp 10–63. But note that their analyses do not result in pure mathematical equations in the consequents. I cannot pursue the significance of this point here

Scientists reason from false assumptions in conjunction with asserted premises to derive the false equations. When they are done with their derivations, the conclusions that they validly draw are conditional in form. This is the way valid reasoning from assumptions works. That the conditionals can be vacuously true should be no surprise. If so, the question becomes: When are the resulting vacuously true generalized conditionals acceptable? My summary answer to this question is given at the end of this article.

III. THE FIRST EIGHT PRINCIPLES

Now, assume that the above analysis for mathematical laws is correct. Then, contrary to many realists, one should not take the ideal gas law to be approximately true. Instead, it is true and known to be so before any testing takes place. Also, contrary to van Fraassen, one should not accept (II) above *because* it is true; nor does it make any sense to accept this *true* generalization because it is *approximately true*. But while it is also clear that one could say that the *equations* of empirical science are false, following Nancy Cartwright,[5] from my perspective it is better to say that calculated values ("*C*-values") obtained through the *laws* do not agree with measured values ("*O*-values") for the same variable.

Here then is the scientific realist position I wish to develop:

(1) The mathematical laws or theories of science are true and known to be so prior to testing them.
(2) The equations of science are false and known to be so prior to testing them.
(3) The mathematical laws are deliberately constructed so as to misrepresent reality, that is, they are not attempts to provide a literally true story.
(4) The laws or theories are true, but they are not accepted because they are true.
(5) There is no role for approximate truth.
(6) The aim of testing is not to falsify; it is to determine how large the

---

[5] *How the Laws of Physics Lie* (New York. Oxford, 1983). At least in this work Cartwright does not take the laws to be in part constituted by the ideal conditions. From my perspective, one obtains her false "laws" by stripping out the idealizing predicates in cases like (II) above.

expected differences are between the values of the predictions and the measurements for the same variable.

(7) It is a major goal of science to justifiably explain the differences.

(8) Being able to explain these differences is the basis for acceptance of laws or theories that generate them.

One may now understand how to take such laws to be true and yet maintain that they misrepresent reality. And here the misrepresentation is deliberate. The scientist is in the position of facing a messy reality, and in order to have a computable function "ignores" certain factors by idealizing them, where these factors are relevant to the values of the variables of interest. In the ideal gas law, the initial variables of interest are P, V, and T. But the scientist knows that the actual shapes of the molecules, their imperfectly elastic collisions, their attractive forces, and their nonnegligible volumes affect the values for P, V, and T. She knows that if one idealizes these factors away, that the *C*-values and the *O*-values will not agree. In fact, the *existence* of these factors explains why the *C*- and *O*-values are not the same. In creating an imaginary world where the mathematics will "work," the scientist does not lose sight of what has been idealized, since it is these idealized factors that are the first court of appeal in explaining the expected differences.[6]

For example, suppose that we use *O*-value inputs for V and T of a gas and calculate the output *C*-value for P. Call the calculated value for P, Pc, and the measured value, Po. The scientist expects that the absolute value of the difference between Pc and Po will not be equal to 0: $|Pc - Po| > 0$. She is now asked to explain why this is the case. She says that as long as we ignore the actual volumes of the molecules that we can expect a difference. The existence of the actual volumes of the molecules *explains in part* why $|Pc - Po| > 0$.

And so it does not seem quite right to say that science aims for a literally true story. In the face of too many factors, difficulty with units and so forth, perhaps one should see the goal to be one of computable functions based on ideal conditions. But if this is so, then why should anyone, Cartwright included,[7] want to ignore the ideal conditions and their role in theorizing? More importantly, how does one know when a true idealized law is acceptable? Instead of the latter question, however, the question normally asked is when are the (false) equations acceptable?

---

[6] They are not the only court of appeal, however.

[7] In *Nature's Capacities and Their Measurement* (New York: Oxford, 1989), Cartwright seems by implication to take the ideal conditions to be constitutive of the laws, since she here (p. 204) proclaims Nowak's method of concretion to be the method of science. Nowak does take the ideal conditions to be constitutive of the laws.

IV. HOW TO EXPLICATE "CLOSE ENOUGH" RESULTS FOR FALSE EQUATIONS?

The false equations are normally taken to be acceptable only if the $|O - C|$ results are "close enough." Certain approximate truth approaches will try to specify when the $|O - C|$ interval is "small enough" so that one is justified in saying that the theory, or law, that entails the C-values in question is approximately true. Of structuralist approaches to specifying such intervals Jeff L. Ramsey[8] says that "when it comes to specifying what counts as an acceptable range of discrepancy, structuralists resort to hand-waving and vague generalities" (*ibid.*, p. 155). I agree with Ramsey, and go further to say that there is no manner of knowing independent of the theory/law at issue when an interval is close enough.

It is instructive to look at Ronald Giere's treatment of the issue of acceptability in the face of $|O - C|$ differences. Instead of "close enough," Giere[9] uses the term 'similar' when he states that the real system must be "*similar* to the proposed model in specified *respects* and to specified *degrees*" (*ibid.*, p. 80). In making this proposal Giere is rejecting van Fraassen's proposal[10] that a theory/model is empirically adequate and hence acceptable only if the empirical substructure of the model is isomorphic with the real system modeled. Giere believes that isomorphism is much too strong a relation to require, and in this I am in agreement. In fact, I am convinced that there are few, if any, empirically adequate mathematical laws or theories on van Fraassen's account because the O- and C-values are never identical.

In his later work, *Explaining Science*,[11] Giere is more insistent than before on the need to recognize the role of idealizations in science. Here he states that "idealization and approximation" are of the essence of empirical science, and that any theory of science "must reflect this fact in its most basic concepts" (*ibid.*, p. 78). While I am in complete agreement with these kinds of remarks, and I am sure that Leszek Nowak, Cartwright, Ramsey, and a handful of others would

[8] "Towards an Expanded Epistemology for Approximations," in D. Hull, M. Forbes, and K. Okruhlik, eds., *PSA 1992: Volume 1* (East Lansing, MI: Philosophy of Science Association, 1992), pp. 154–64.

[9] "Constructive Realism," in Paul M. Churchland and Clifford A. Hooker, eds., *Images of Science: Essays on Realism and Empiricism, with a Reply from Bas C. van Fraassen* (Chicago: University Press, 1985), pp. 75–98.

[10] See van Fraassen, pp. 44–46.

[11] Chicago: University Press, 1988.

also agree, it remains to be determined whether Giere's criteria for acceptability in the face of the $|O - C|$ differences generated by idealizing in theory construction are adequate.

In *Explaining Science*, Giere continues with his view that a model/theory is acceptable only if the model is similar in certain respects and to certain degrees to the real system. He illustrates his approach using the Earth and Moon as the real system, taking the model to be that of a two-particle Newtonian system with an inverse square central force. Here he says that the respects are "position" and "velocity," while the degree is claimed to be "very close" (*ibid.*, p. 81). It seems fair to say that Giere means by "respect" what I have called "variables of interest," and, at least in this example, these variables have units. His degrees then have to do with how close the C-values of the variables of interest from the model are to the O-values of the same variables for the real system. If this is the correct read of Giere's position, then I have problems with both his "respects" and his "degrees."

When idealizations are invoked, the real system and its model are always going to be dissimilar in *relevant* respects. And the respects where the dissimilarity holds may not be representable with variables of interest with units. In the above example, the model consists of two-point masses with only gravitational forces acting between the two bodies and no other forces acting on the bodies. The real system has no point masses; it does not even have perfectly spherical, rigid masses. There are other forces operating on both the Moon and the Earth. As is well known, these are extremely relevant dissimilarities. In fact, these dissimilarities reflect what has been idealized and how. Any assessment of the acceptability of a model or theory that fails to take account of relevant dissimilarities between the model and the real system is going to be inadequate. In what follows I hope to show why such an account, and hence why Giere's account, is not adequate when it comes to "respects." In so doing I hope to show why some dissimilarities are relevant and others are not.

As far as Giere's degrees are concerned, I do not see how he gets past Ramsey's criticism of the structuralists. On what grounds does he make the claim for the above example that the $|O - C|$ differences are "very close"? Why are these differences not "great" instead of "very close"? Or why not "sort of close" instead of "very close"? Perhaps Giere has answers for these questions, but I do not know what they are. In what follows, I will argue for an answer for when one is justified in saying that the $|O - C|$ differences are "close enough."

V. A DIFFERENT ACCOUNT OF "CLOSE ENOUGH"

An engineer might take some $|O - C|$ differences to be close enough for some engineering goal, and nevertheless not be the least bit interested from a theoretical perspective why he has these differences. The sense of "close enough" in which I am interested here is the theoretical one. The history and practice of science is full of cases where the scientist is interested in why he has the $|O - C|$ differences that he does. Being able to successfully explain these differences will require that there exist factors that generate or are productive of these differences. It is when the scientist thinks he knows what the factors are that are productive of the differences, that he will take the differences to be close enough from a theoretical perspective. And so from an epistemic perspective we have the following explication for the theoretical sense of "close enough":

> (A) For any *O*- and *C*-value pair, an $|O - C|$ difference is "close enough" for a scientist if and only if he can provide a complete and epistemically justified explanation for why $|O - C| > 0$.

Having rejected truth à la van Fraassen or approximate truth as a reason for accepting an idealized law or theory, we now have through (A) above a necessary condition for acceptance:

> (B) A true idealized law or theory is a good one, and is hence acceptable only if its $|O - C|$ differences can be completely and justifiably explained.

In support of (B), the history of science surely demonstrates that if one cannot completely and justifiably explain the differences, then the law or theory is in trouble. And so philosophers of science, in the grip perhaps of the deductive nomological conception of explanation, have failed to recognize one of the more significant facts in need of explaining, namely, the differences between the *O*- and *C*-values (hence, the subtitle of this essay). But this is not all that has been missed. There are other kinds of differences that are in need of explanation. Again, the history of science is full of cases where one has incompatible *C*-values for the same variable generated by different "methods," for example, incompatible diameters for a Nitrogen molecule. The scientist is uncomfortable unless he can account for such differences. But perhaps more importantly, there are plenty of cases of incompatible *C*-values for the same variable, and the scientist thinks that he can explain why one of the *C*-values is more accurate than the other. How do such explanations work? We will have more to say later about these matters, but for now we have three candidates for explananda that have been basically neglected by philosophers: $|O - C|$ differences; incompatible *C*-value differences for the same variable obtained using different "methods"; the greater accuracy of one *C*-value over another for the same variable.

VI. HOW TO JUSTIFIABLY EXPLAIN THE $|O - C|$ DIFFERENCES

We now examine the type of explanation in question and four "methods" whereby one can provide epistemic justification for claiming to have explained the differences in whole or in part. There are in fact at least five "methods" if one includes the use of measurement or experimental error. The ideal gas law falsely assumes that the values for the molecular volumes of a gas do not affect the values for V, P, or T. Assume that we have a C-value for pressure of Pc and an O-value of Po, where $|Po - Pc| > 0$. Why is there this difference between these two values? In effect, it was Rudolph von Clausius who said that it was because we have neglected, among other factors, the actual volumes of the molecules in the container with our idealization about negligible volumes. Clausius reasoned that the volume occupied by a gas should not be taken to be the volume of the container V, but that it should be the volume available to a single molecule, given the volume taken up by all the molecules in the container. The volume available is the volume taken up by all the molecules, which is b, subtracted from V. The amount subtracted from V, that is, b, is a function of the number of molecules in the container and their radii. Clausius's corrected version of the ideal gas equation is: $P(V - b) = RT$. In correcting the original gas equation, Clausius drops the gas law's false assumption that the molecules have negligible volumes. With the corrected equation, we have a new C-value for pressure, where this new value has *moved closer* to the O-value, that is, we have increased the accuracy of our predictions. This increase in accuracy provides an epistemic justification for our belief that the actual volumes of the molecules are factors affecting the values of our variables. We now claim to know that if we ignore the actual volumes and number of the molecules in the container, we will get deduced results that are different from our measured results. Put another way, we now are justified in claiming to know that the number and volumes of the molecules in the container are causal factors that make gases have the measured values that they do. We are justified in claiming that we have explained *in part* why we have a difference in C- and O-values for the ideal gas law.

In the example of Clausius's equation, one does not have a *complete* explanation for the $|O - C|$'s for the original ideal gas law by appeal to the actual volume taken up by the molecules. But one knows that one has explained in part these differences, since one has made the C-values move closer to the O-values in correcting the original equation. This means that the $|O - C|$'s are still not close enough. Meanwhile, J. D. van der Waals drops the ideal condition that the molecules exert no attractive forces on each other and corrects

Clausius's equation by factoring in a correction for pressure based in part on the number of molecules in the container. Once again, the C-values move closer to the O-values, but also once again the $|O - C|$'s have not shrunk to zero. And so, one has explained in part the differences, but the differences are still not close enough. The differences are not close enough until they have been completely and justifiably explained. Of course, it is *possible* that one may eventually drop all recognized ideal conditions and correct for them so that the $|O - C|$'s *are explainable* by appeal to measurement error.[12] While this is a definite possibility, I have yet to come across an equation that is the end result of this process of dropping and correcting that has shed all of its idealizations. What this means is that we need other ways to justify the claim that an idealized factor explains the differences. A complete explanation of the differences requires that one identify all the factors that affect the outcome. Dropping and correcting is but one manner to justify a claim that a factor is explanatory.

Now briefly consider some additional cases of unexplained differences: Why were the O- and C-values for the position of Uranus different? Neither the idealized factors for Newton's laws nor measurement error could account for the differences. It was the existence of Neptune that ultimately did the explaining. Or again, why did the O- and C-values differ for the perihelion of Mercury? No factor was found to explain the difference, and so the true idealized Newtonian laws are in trouble. But one can correct for Newton's constant mass m—that is, instead of m, one writes $m/(1 - v^2/c^2)^{1/2}$ (thereby creating the Special Theory of Relativity). This gets our C's to move closer to our O's, for example, in the Mercury case. In doing this, we recognize that the original uncorrected laws require an additional idealizing assumption, namely, that mass is a constant. Or consider the Davisson/Germer experiment to determine the wave length of a 54-volt electron, using a Nickel crystal and Bragg's Law to get one value, and the De Broglie equation to get the other. The difference was only 0.02×10^{-8} cms. between the two values! This was not "close enough" for them, and they explained the difference ultimately in terms of refractive indices for electrons, correcting Bragg's Law in the process.[13]

---

[12] See Sears, chapter 11 for the gas law account

[13] See J. D. Stranathan, *The "Particles" of Modern Physics* (New York: McGraw-Hill, 1942), pp. 541–45, for an account of the Davisson/Germer experiment Note that this example is not really a case of an O- and C-value difference, but rather it is a case of a difference between two incompatible C-values of the same variable. Here there is no independent O-value to get by measuring or observation See section xiii of this paper.

VII. THE FIRST METHOD OF JUSTIFYING AND NOWAK'S WORK

This method of dropping and correcting is Nowak's method of strict concretization. At least in his 1972 and 1980 works (*op. cit.*), Nowak takes the idealized laws to be vacuously true, and such laws are acceptable if they are elements in a sequence of dropping and correcting with increasing accuracy. Leaving aside initial conditions and auxiliary assumptions Nowak's sequences look as follows, where T_n is a law (universal conditional), I_n an ideal condition, and C_n a *C*-value:

(a) $[(T_1 \,\&\, I_1 \,\&\, I_2 \,\&\, \ldots\, I_n) \rightarrow C_1]$
(b) $[(T_2 \,\&\, I_2 \,\&\, \ldots\, I_n) \rightarrow C_2]$
(c) $[(T_n \,\&\, I_n) \rightarrow C_n]$

Here Nowak is dropping an ideal condition and correcting the original "theory" with the hoped for new and more accurate *C*-value. The I_n are singular ideal condition statements as already described. The ideal conditions are also "built into" the T_n as unsatisfied predicates in the antecedents. For Nowak, if for each dropping and correcting episode one gets increased accuracy, then the sequence and each T_n in the sequence is acceptable. But Nowak recognizes that some of the I_n will not be amenable to drops and corrections. At this point Nowak introduces his method of *approximate* concretization. This method requires one to specify a threshold interval "in advance" such that when the $|O - C|$ differences fall within this interval, one can say that the results are close enough. Here he joins Ramsey's hand-wavers.

Nowak does not see his method of strict concretization as a method of epistemically justifying the explanation for why $|O - C| > 0$. He thus does not see that explaining the differences is a major goal of science. For me, his method of strict concretization is but one method out of at least five for providing epistemic justification. To be sure this point is clear, consider the following schematic example. To keep things manageable, we will concentrate upon one mathematical law. Let us assume that we believe that a given magnitude F_1 is in some way a function of the factors F_2 through F_7. We see that coming up with a computable function will be very difficult, given so many variables and their units. So we idealize some of the factors, say F_4 through F_7. Let I_n be the idealizing condition for F_n. As a first shot, our mathematical law becomes:

(III) If $I_4 \,\&\, I_5 \,\&\, I_6 \,\&\, I_7$, then $F_1 = f(F_2, F_3)$

Here (III) is just the general structure for (I) in the above for the ideal gas law. In the best of circumstances, the F_n will have units, but this is not always possible or easily accomplished. From my perspective, the *premodeled* real system must be characterized in some manner. My way of characterizing the real system for mathematical laws is to speak of noncomputable functions. And so we may say that F_1 is a noncomputable function of F_2 through F_7. To say that F_1 is a noncomputable function of the other factors is to say that we can bring about increases or decreases in F_1 by changing some of the other factors in some manner. The various ways in which such a claim can be checked may involve, not surprisingly, J. S. Mill's method of concomitant variation. The important point to note here, however, is that the real system is characterized by saying that *there exist factors* F_1 *through* F_7 such that F_1 is a noncomputable function of F_2 through F_7. Instead of a noncomputable function, Nowak talks about the image of a law or theory, where the latter is a kind of list of what he calls "primary and secondary" factors.[14] The secondary factors are the ones initially idealized.

Now Nowak will accept (III) above if he can first drop, say, I_4 and correct for it with the result:

(IV) If I_5 & I_6 & I_7, then $F_1 = f(F_2, F_3, F_4)$.

And (IV) must now be more accurate than (III). And (IV) must also be acceptable, which requires greater accuracy in:

(V) If I_6 & I_7, then $F_1 = f(F_2, F_3, F_4, F_5)$.

Assume now that we are unable to drop and correct for I_6 and I_7. At this point Nowak resorts to the $|O - C|$'s falling within some specified interval in order to provide acceptance for the set of laws. This is a part of his position that I cannot accept. For me, the above process should be seen differently. One may best see how I interpret the process by pondering the following question: When is an ideal condition relevant? My answer is that a given idealizing condition I_n is relevant only if the factor F_n that it idealizes exists and explains (at least in part) why $|O - C| > 0$.[15] From my perspective then, dropping and correcting is but one method for providing epistemic justification for the claim that a given F_n exists and explains the $|O - C|$'s. It is a method of establishing that an ideal condition is relevant. We can thus state a necessary condition for the acceptance of an idealized mathematical law. Such a law is acceptable only if all of its explicitly recognized ideal conditions are relevant. And again, such ideal conditions are relevant only if the factors they idealize exist and play the appropriate explanatory role.

---

[14] *The Structure of Idealization*, pp. 117–20. I came upon the need for "lists" or noncomputable functions independently of Nowak. Cartwright also believes, possibly under Nowak's influence, that the factors omitted (and presumably the ones that are not) are on a list. But for individual cases she believes that this list is never completed—see her *Nature's Capacities and Their Measurement*, p. 207. For me, the issue of a complete list opens up a whole new set of questions

[15] Hereafter I will leave off the modifier 'at least in part' as understood.

VIII. TWO MORE METHODS OF JUSTIFICATION

There are at least two other methods involving ideal conditions whereby one may epistemically justify an explanation for the $|O - C|$'s. First, one may make one's experimental conditions more closely resemble or approximate the ideal conditions with the hoped for result *that one's O-values move closer to one's C-values.* Here the equation remains the same without any added corrections. Here manipulating a factor F_n relative to its I_n so as to make F_n more closely resemble I_n—with the happy result that the gap between the O- and C-values narrows—justifies one in claiming to know that F_n exists and explains why $|O - C| > 0$. Second, one can drop one ideal condition and correct for it, but in the process one may introduce a new ideal condition. Increased accuracy, however, justifies one in thinking that the F_n idealized exists and explains the differences.

An example is helpful here. Abstractly represented, an Atwood machine consists of a pulley over which is run a cord to both ends of which are attached equal masses M. When a rider mass m is placed on one of the larger masses M, the acceleration of m is given by the following equation: $a = mg/(2M+m)$, where g is the acceleration of gravity. The false assumptions that are invoked here are at least the following:

(A_1) The pulley has no mass.
(A_2) The cord has no mass.
(A_3) The cord has unchanging length.
(A_4) There is no air friction.
(A_5) There is no friction on the pulley's axle.

For convenience we avoid here the generalized conditional for this mathematical law, that is, we avoid the analogue to (2) above for the ideal gas law. Instead, we use a schema as before:

(VI) If A_1 & A_2 & A_3 & A_4 & A_5, then $[a = mg/(2M+m)]$.

Assume that we have a big, heavy iron pulley. As a result, our contraption barely moves when m is placed on one side of the pulley. We, of course, believe that the mass of the pulley affects the differences between our O- and C-values for acceleration. We claim to explain why $|O - C| > 0$ by citing the fact that the mass of the pulley is not equal to 0. We are now called upon to justify this claim. We do so by replacing the iron pulley with a lightweight aluminum one. As a result, with the same values for *M*, *m*, and *g*, we get our *O*-value for acceleration to move closer to the *C*-value for the same variable. This movement justifies our explanatory claim that the mass of the pulley exists and is a factor that affects our value for acceleration. Making an F_n more closely resemble its I_n with a movement of the *O*-values closer to the *C*-values is another method to justify one's explanatory claims, another method to establish the relevance of an I_n in a given law. But note that in this method the ideal condition is not dropped and the equation is not altered.

The other method consists in dropping an I_n and correcting for it, but in such a way that one replaces the original I_n with a second ideal condition. For example, in the Atwood machine equation we can introduce the quantity of the mass of the pulley, but to do so we must idealize the configuration of the pulley's mass. We thus drop (A_1) and replace it with (A_1'). If our *C*-values for acceleration move closer to our *O*-values, then again we are justified in believing that the mass of the pulley explains the differences. But we will presumably never completely close the gap between the *O*- and *C*-values for acceleration as long as we have an equation constructed for a mass distribution that the pulley does not have.

And so to completely explain the $|O - C|$ differences, we will have to have all the factors that affect the outcome on some kind of list. The three methods just discussed show how to justify a claim that a factor is relevant and also show how to determine a necessary condition for the relevance of an ideal condition. In the last two methods, we end up with laws that are still vacuously true. In the first method, we might be able to get a nonvacuously true law, but given the number of variables that would be required and the issue of computability, and given the fact that some factors cannot be captured with variables, this seems unlikely. It suffices to say that large numbers of laws cannot be rid of their idealizations.

IX. FURTHER COMPARISON WITH NOWAK

Nowak seems not to recognize the possibility of dropping one I_n with a correction and replacing this I_n with a new ideal condition. This is due in part to his requirement that the factors idealized have units, something that is not always possible as the pulley's reconfigured mass illustrates. By comparison to his ideal conditions, my ideal conditions are probably best construed as "false assumptions." Nowak also seems not to recognize the role of making the F_n more closely resemble their I_n, although he does talk about the possibility of conducting perfect experiments where one gets his F_n to equal his I_n.[16] My explanation for his failure to recognize the role of making the F_n move toward the I_n is that he does not see the process as one of explaining the differences and justifying these explanations by increasing accuracy. To see this, ponder what Nowak would do with a case with seven ideal conditions for one equation, where one cannot correct for a single one of these conditions so as to have a new equation. To accept such a law, he would have to be able to say for each I_n that it is "close enough" to its F_n such that the $|O - C|$'s are "close enough." Now this is exactly what I do not think anyone is able to do. On the other hand, on my approach one can maintain that the differences between the O- and C-values are generated by the factors F_1 through F_7 and one can do this by attempting to make our F_n move closer toward the I_n with the happy result of a narrowing of our $|O - C|$'s. To be sure, we may be mistaken about whether we have all of the factors that affect the differences, but, if we do have all of the factors and can for each of them decrease the gap,[17] then we are justified in our claim to have completely explained the differences. Nowak seems not to see that if there really are only F_n idealized relevant factors and we show for each of them, without correcting, a decrease in $|O - C|$, then we have a complete explanation[18] for why $|O - C| > 0$.

As an example of this kind of case, consider the Hardy-Weinberg Law. This law tells us under what conditions the gene ratios and genotype frequencies will remain constant for a natural population. Ignoring the equation, the ideal assumptions are: there is an absence of recurring mutations, an absence of any degree of preferential matings, an absence of differential mortality or fertility, an absence of immigration or emigration of individuals, and an absence of fluctuations of gene frequencies due to sheer chance.[19] This law seems to hold under at least seven idealizing assumptions, where, as far as I know, one cannot correct for any of these assumptions so as to obtain greater accuracy in general. But no one would want to say that the idealized factors do not affect the values of the variables. Both Nowak and I want to drop and correct with increasing accuracy. I believe that this is but one way to justifiably explain the differences, however. Nowak does not see this as providing justified explanations. When we both meet an I_n that cannot be dropped, Nowak resorts to what Ramsey[20] calls "hand-waving," while I attempt to show that an idealized

---

[16] See *The Structure of Idealization*, pp. 111–12.

[17] We will discuss the complication of "gap increase" below.

[18] Such claims to completeness include measurement errors as factors

[19] See Alexander Rosenberg, *The Structure of Biological Science* (New York: Cambridge, 1985), pp. 132–33.

[20] I note in passing that Ramsey also does not see the goal as one of explaining

F_n affects the outcome by changing the F_n to make the O-values move closer to the C-values.[21]

X. ONE MORE METHOD OF JUSTIFICATION

There are numerous cases in the history of science in which the differences cannot be explained by appeal to some explicitly idealized factor or measurement error. The discovery of Neptune to explain the differences for Uranus is a case in point. The hypothesis chosen to explain the difference in the Uranus case is an existential hypothesis: there is some planetary mass other than that of Earth, Mars, Venus, and so forth. This existential hypothesis is proposed because the ideal conditions fail to mark factors that adequately explain the differences. This existential is *strictly verified* by finding a single, individual, particular, namely, Neptune. This example illustrates yet another manner whereby one may justifiably explain the differences. *When the F_n idealized by the I_n fail to explain the differences, then one looks for factors heretofore not recognized.* It should be clear, as it was for this case, that the search for factors not idealized can be guided by the noncomputable functions or lists for the accepted equations. The lists are statements of relevance.

---

the differences, although he does see the need to epistemically justify what he calls the "approximations," that is, my C- and O-value differences

[21] Lest some may think either Nowak's or my position are akin to Ronald Laymon's position in his "Idealizations and the Testing of Theories by Experimentation," in Peter Achinstein and Owen Hannaway, eds., *Observation, Experiment, and Hypothesis in Modern Physical Science* (Cambridge: MIT, 1985), pp. 147–73, one should note my formalization of Laymon's sequences: (a) $[(T \,\&\, I_1 \,\&\, I_2 \,\&\ldots I_n) \rightarrow C_1]$; (b) $[(T \,\&\, I_1' \,\&\, I_2 \,\&\ldots I_n) \rightarrow C_2]$; (c) $[(T \,\&\, I_1' \,\&\, I_2' \,\&\ldots I_n') \rightarrow C_n$. The important thing to notice is that T remains constant in his sequences and is not known to be (vacuously) true prior to testing. Increasing accuracy in the sequence confirms that the unchanging T is true; while failure to increase accuracy confirms that T is false. Note also that Laymon only talks about replacing ideal conditions with more realistic ones and he thus does not talk about dropping them without replacement. And so I_1 goes to I_1' for Laymon, while for Nowak, I_1 is simply dropped. Laymon also does not *explicitly require* that a correction be made when he replaces one ideal condition with a more realistic one. Thus, just about any T can be disconfirmed by moving from, for example, 'the molecules are perfectly spherical' to 'the molecules are nearly spherical', where without any corrections we do not get greater accuracy. Also, Laymon repeatedly takes an equation or conjunction of equations to be a theory. Now either these equations are asserted conditionally on the basis of their ideal conditions or they are not. If they are, then the result is true If they are not, then they are false and known to be so prior to testing. Thus, Laymon misconceives the process as one where we need to find out whether a given T is true or false Space forbids my arguing these points here, but hopefully one understands why I maintain that Laymon's position is very different from that of Nowak (and also my position) and that what he and others need to recognize is that the ideal conditions are constitutive of the laws and make them true, so that the goal is to explain the differences. See also Laymon, "Cartwright and the Lying Laws of Physics," this JOURNAL, LXXXVI, 7 (July 1989): 353–72.

XI. A FICTIONALIZED ACCOUNT OF THEORIZING

Here is a fictionalized schematic account of the elements of theorizing using the ideal gas law. Assume that Mary believes that the pressure of a gas is affected by its volume and temperature and that she has units for these magnitudes. She also believes that the volumes of the molecules (F_1), attractive forces between them (F_2), shapes of the molecules (F_3), and imperfectly elastic collisions (F_4) are also factors that affect the values of P, V, and T. She tries to construct as accurate a computable function as possible by idealizing the factors other than P, V, and T. In a sense she is maintaining that P is a noncomputable function of V, T, F_1, F_2, F_3, and F_4. She idealizes F_1 through F_4 to get the following:

(VII) If F_1 & F_2 & F_3 & F_4, then $P = f(V, T)$.

Mary takes (VII) to be true and also reasonably expects the O- and C-values to be different. She knows that if F_1 through F_4 are all of the factors that affect the $|O - C|$ differences (outside of measurement error), then she should be able to justifiably and *completely* explain the differences. She sees that her ideal conditions are relevant only if the factors they idealize exist and explain the differences at least in part.

What this fictional account reveals is an element of theorizing that seems to be missing in this day of treating theories as models of real systems. What is missing is the characterization of the real system modeled. In the above, this characterization is given in terms of a noncomputable function or list. If one takes such "lists" to be elements in theory construction, then one will want lists that have all and only relevant factors on them. The statement describing such a list will not be taken to be approximately true. Instead, it may be taken to be false because incomplete. Or it may be considered false because it has factors on it that are irrelevant or do not exist. Importantly, an underlying element of theorizing for a mathematical law or theory may be seen to be a list or a statement describing the list. The latter, as already indicated, is a kind of statement of relevance. We will want a "literally true story" for this list and we will accept the statement describing the list in part *because* it is true. For example, no matter what equations one might set forth, phlogiston theory is false because there is no phlogiston; caloric theory is false because there is no caloric substance; Ptolemaic theory is false because there are no epicycles.[22]

---

[22] I take my characterization of theories here to be an explication of what Ian Hacking has called "entity realism"—see his "Experimentation and Scientific Realism," *Philosophical Topics*, Volume XIII (Fayetteville: Arkansas UP, 1983), pp. 71–87

XII. VAN FRAASSEN'S ACCIDENTAL REALISM AND TWO FINAL PRINCIPLES

van Frassen does not take such lists or their characterizations to be elements of theorizing. Nor, for that matter, do many realists.[23] van Frassen's portrayal of realism "accidentally" works for lists; it does not fit the idealized laws constructed from such lists. To the eight principles given above characterizing my kind of scientific realism, we now add:

(9) The mathematical laws are constructed on the basis of lists of factors taken to be relevant to each other.
(10) An ideal condition I_n for such laws is taken to be relevant only if the factor F_n that it idealizes exists and plays the appropriate explanatory role.

These ten conditions do truly provide a different conception of scientific realism, without appealing to "approximate truth"; they show that science does and should aim at explaining the differences.[24]

XIII. THE ADDITIONAL EXPLANANDA

The Davisson/Germer example mentioned above involving electron wave lengths shows that there are cases where the explanandum is not $|O - C| > 0$. One should not take the wave length value calculated using Bragg's Law as an O-value (as I did in the above). Instead, one should see that this is a case of *incompatible C-values for the same variable.* These cases abound in the history of science. These are cases where one cannot get one's hands on a value except by calculation. There is no "independent" measurement or observation that gives one an O-value to compare with a C-value. Let us represent the cases of incompatible C-values for the same variable as follows: $|C_1 - C_2|$. Now there are two problems that arise here, where the first problem is the same as before—namely, what is the explanation for the difference? The second problem is given by the following question: *Which of the incompatible C-values obtained by different methods is the most accurate?* The issues connected with these questions require another article. These are extremely important issues in the practice of science, but they have received little attention from philosophers. The scientist attempts to answer these questions by appeal to the ideal conditions invoked for each method, as well as some of the other elements already mentioned.

---

[23] Nowak and now maybe Cartwright, in *Nature's Capacities and Their Measurement,* are exceptions.

[24] The realist has a right to expect some account of the role of ideal conditions from the antirealists. When are such conditions relevant? For the ideal gas law, why not assume the molecules to be perfect cubes or why not assume that gases contain tribes of tiny invisible elves at war?

Implicit in the problems generated by different *C*-values for the same variable are additional challenges for the antirealists. Briefly: How is the antirealist going to make sense of the practice of scientists selecting one from a set of incompatible *C*-values as the most accurate? For example, the value for electronic charge calculated using the ruled grating method is taken to be more accurate than that obtained by the Millikan Oil Drop Experiment.[25] How is the antirealist going to explain why $|C_1 - C_2| > 0$? How is the antirealist going to avoid having to admit that science is inconsistent, for example, how to "explain away" the incompatible *C*-values for the diameter of the Nitrogen molecule calculated using three different methods?[26]

What we have then are three neglected explananda, where all three are inextricably tangled up with idealizations and the existing factors idealized. From my perspective, the most interesting of the three involves explaining the greater accuracy of a *C*-value without appeal to measurement or experimental error. Surely, this greater accuracy is explained in at least some cases by appeal to the fact that there are (observable or unobservable) factors that affect the outcome and fewer of these are idealized in one method by comparison to another.

XIV. COVERING LAW EXPLANATIONS AND CARTWRIGHT'S PROBLEM

If the equations are false, then one cannot explain using the covering law model of explanation. There are a number of reasons why this is so, but I focus only on the fact that one cannot explain why an *O*-value obtains by deducing it from some true initial conditions and the covering laws. This is so because one deduces a "false" *C*-value instead of the true *O*-value. Cartwright tries to deal with this problem of explaining true *O*-statements by deducing false *C*-statements by offering us her simulacrum account of explanation.[27] It is not necessary to go into the details of her simulacrum account here, since she herself thinks that it is not adequate. Instead, she believes that what is needed is "a theory of explanation which shows the relationship between causal processes and the fundamental laws we use to study them, and neither my simulacrum account nor the traditional covering-law account are of much help" (*ibid.*, p. 162). It is my view that this paper offers such a theory of explanation, one which requires, happily, that the explanandum be what scientists themselves think is in need of explaining. I am thinking here of the $|O - C|$ differences generated by the false equations (her "fundamental laws") being explained by the idealized factors (her "causal processes"). From my perspective we explain these differences without appeal to universal laws.

---

[25] Stranathan, pp. 56–63.

[26] See W. J. Moore, *Physical Chemistry*, Second Edition (Englewood Cliffs, NJ: Prentice-Hall, 1955), pp. 178–80.

[27] *How the Laws of Physics Lie*, pp. 151–62.

To see this latter point, try to find a universal law in the schematic explanation of an $|O - C|$ difference for an idealized law L in the following: Why is this $|O - C|$, generated from ideal law L and input O-values $O_1,\ldots, O_n$ not equal to 0? Because there exists at least one factor F_n whether on or off the list for L, that has been ignored or idealized that affects the values of the target variables for L. The explanatory statement here is not a covering law and seems to be existential in form, where some of these existentials have even been strictly verified. Oddly enough, these existentials seem to require *mention* of the ideal law used to get the C-value, and as such, are partly *metalinguistic* claims. It appears to me that any effort to obtain a universal covering law-like statement to explain the difference will also have to mention the ideal law from which the particular C-value is derived. This is but one reason why I believe that there can be no covering law explanation for such differences, unless one is prepared to admit that there are laws that *mention* laws.

XV. THE PROBLEM OF GAP INCREASE

In the above (footnote 17), I mentioned the problem of "gap increase." This is an interesting problem because it generates a tension between the goals of increased accuracy and having a complete list for a given theory. Assume that one has a list consisting of F_1 through F_5, where F_1 and F_2 are idealized for the function constructed on the basis of this list. Now assume that there are two factors, F_6 and F_7, that are not on the list and are such that the increase caused by F_6 is exactly equal to the decrease caused by F_7. In scenario one, neither F_6 nor F_7 is detected, and so one could get "close enough results" as a matter of luck. In scenario two, F_6 is detected and F_7 is not, and F_6 increases the $|O - C|$ differences when corrected for. In case three, the situation is reversed where we get increased accuracy by failing to have F_6 on the list, while correcting for F_7.

On my approach for case one, the idealized law is not acceptable because the list is incomplete, that is, there are factors not on the list that affect the values of the variables. In the above, I gave a necessary condition for law acceptance (as opposed to "gap" acceptance), namely, that we accept idealized laws only if the idealized factors exist and play an explanatory role. Case one seems to suggest that law acceptance also requires that the list be complete. Case two shows the tension between accuracy and completed lists. Here we can assume that the discovery of F_6 leads to the belief that the $|O - C|$ differences are explained by F_1, F_2, and F_6. Assume that we can correct for F_6 expecting a decrease in our differences which would justify our belief in F_6's existence as an explainer. But we get an increase. What this means from my perspective is that the original difference was explained by F_1 and F_2, and that we now do not know how to explain the increase due to our correction. The causal influence of F_6 has been there all along. Correcting for its absence in the equation seems to entail that there is a canceling factor, for example, F_7. Note that this latter is an existential claim that can be verified. We explain the increase in this case by appealing to the existence of the canceling F_7. In case three, we simply reverse the scenario for case two. But here in correcting for F_7 we get a decrease in our original differences. This would lead one to believe that F_1, F_2, and F_7 explain the differences, but this would be incorrect. As before, F_1 and F_2 explain the original differences and we will be lucky if we discover the canceling effect of F_6. For without this discovery we do not have an explanation for the decrease due to correcting for F_7. Cases two and three also seem to suggest that our idealized law is acceptable only if the list is complete.

Now consider the three scenarios from the point of view of an antirealist or instrumentalist. If these protagonists take one equation to be more acceptable than a second when the first is more accurate than the second, then they are presumably committed to wanting the equation from case three above. If increasing accuracy is *all* that one requires for purposes of acceptance, then our protagonists seem to be committed to ignorance. They do not want to know that there exists a factor, for example, F_6, that explains the decrease due to correcting for F_7 such that the correction for both factors leads to greater inaccuracy by comparison to correcting for F_7 alone. This is the tension between accuracy and a completed list, more poetically, between messy metaphysics and useful mathematics.

For those who wish to take increasing accuracy of the equations as a basis for acceptance or confirmation, as do Nowak and also Laymon, there are problems here for them as well. For such writers case two leads to rejection or disconfirmation, while case three leads to acceptance or confirmation. Looking at case three, we can think of it as one where we drop the ideal condition for F_7 and correct with the result of increased accuracy. This is supposedly sufficient for confirmation or acceptance. But clearly this is not the case when there are canceling factors like F_6. Nor is increased accuracy necessary for acceptance. This is because the contrapositive of the latter would make decreased accuracy sufficient for rejection or disconfirmation, something shown to be false by case two. All I wish to suggest here is that failure to take account of canceling factors will create problems for such approaches.[28]

XVI. WHAT ARE EMPIRICAL MATHEMATICAL THEORIES?

The "picture" that emerges is that empirical mathematical theories should be seen as conjunctions of idealized laws with list statements. The latter tell us in some manner which factors affect the values of a given variable that has units, for example, acceleration. The factors are given in their broadest terms. And so Newtonian acceleration is affected by quantity of mass, configuration of mass, rigidity of mass, and the different kinds of forces. Newton's general force equation and his gravitational force equation each require at the very least that the masses be point masses, collections of point masses, or rigid bodies, which is why both laws constructed from the equations are vacuous truths. One is not required to incorporate, for instance, "massless pulley," into either law; nor any other more specific idealization which one encounters in the myriad uses of these laws. We have the two Newtonian laws and the list as a basis for deriving other laws, for example, the Atwood Machine Law, where in this instance it is the list that tells us that the cord's mass and elasticity are factors to idealize; that frictional forces need to be recognized, and so on.

In the Uranus and Mercury examples, it is the Newtonian list that tells one to look for another mass as an explainer of the gap. This yielded success for Uranus, but left us without a satisfactory explanation of the gap for Mercury. If we take the Newtonian list to be in statement form, then it might read: these are all and only the factors for acceleration. This statement historically turns out to be false, but it can be corrected by the addition of the inconstancy of mass factor. For me, it is in this sense that Newtonian theory is shown to be false, that is, it is the list statement that is false. The laws are true, but are unacceptable unless corrected or further idealized. This way of looking at matters explains why we take phlogiston theory, caloric theory, or Ptolemaic theory to be false. For all of these theories there were alleged explanatory factors on the lists that did not exist. And so there is no phlogiston or no phlogiston with negative weight, and so on. In Newton's case there was a factor missing from the list.

---

[28] Ramsey recognizes the gap increase problem in his own way.

XVII. WHEN TO ACCEPT A VACUOUSLY TRUE MATHEMATICAL LAW

Finally, I have said that I am looking for a set of conditions that can be used to pick out the "good" from the "bad" vacuous truths. Ignoring the role of measurement error, these conditions seem to be the following for idealized law (L):

(L) If (I_4 & I_5 & . . . I_n), then $F_1 = f(F_2, F_3)$.
(1) $|O - C| > 0$.
(2) Each I_n idealizes an existing F_n on L's list.
(3) Each F_n explains in part why $|O - C| > 0$.
(4) The F_n on L's list are all and only the F_n that affect the values of the variables for L.
(5) Each I_n is required for the derivation or construction of L.
(6) $f(F_2, F_3)$ is the best fitting function relative to the O-values for the unidealized variables for L.

Something is fishy if we idealize and do not get gaps; hence (1). An I_n is relevant only if it idealizes an existing F_n on L's list and the latter plays an explanatory role; hence (2) and (3). We need to watch out for unknown factors and canceling factors; hence, (4). With (5) we avoid perfect cubes (as opposed to perfect spheres) and strange equations like, for example, [$PV = RT^4$]. Gaps are acceptable if and only if completely explained; ideal laws are acceptable only if the gaps are acceptable, and hence only if (1) through (5) obtain for L. We will want the greatest accuracy possible relative to the O-values of the variables that are actually factored into L; hence (6). Now, while I am persuaded that the above six conditions are necessary for specifying when one has a good vacuously true mathematical law, and hence are necessary for the acceptance of such a law, it is not clear whether they are conjointly sufficient. For example, what role should explaining the other two explananda play in acceptance? Even so, the set of vacuous mathematical truths is considerably narrowed in terms of acceptance by the above. These vacuously true laws are not trivial.

XVIII. FINAL DELIBERATIONS

First, let me re-emphasize the fact that this paper is not concerned with laws of a nonmathematical type. There may be such laws of the type 'All X's are Y's', for example, "All Cu wires are good conductors of electricity," and they may even be exceptionless. There may even be laws of this form that have exceptions such that one may nevertheless wish to say that they are approximately true. Such laws may come with provisos[29] or *ceteris paribus* clauses. But such laws are not the concern of this article. The mathematical laws of concern in this article are those that in some manner consist of false equations derived or constructed on the basis of false assumptions. These laws have *C*-value predictions that are not expected to agree with the *O*-values of the real system. For this type of idealized law there are no positive instances, or, if there are, they are the exceptions and are cause for concern. (How to get positive instances if the ideal conditions are relevant?) For laws of the form 'All *X*'s are *Y*'s', the positive instances are the rule, and the negative instances are the exception and cause for alarm.

---

[29] It is remarkable from my perspective that when Carl Hempel talks about provisos

Second, some are going to insist that in some instances the *C*-values are so close to the *O*-values that it is silly to forbid calling the law that generates the *C*-values "approximately true." Here I think the question is: What do we take the law to be in the first place? I have argued in this article that the ideal conditions or false assumptions are constitutive of the laws. I have also chosen to stay with the material conditional. Given my analysis, it is just plain false to say that such laws, as I have presented them, are approximately true.

For those who, at least for the sake of argument, grant that the laws are true, but insist that the equations are false and can legitimately be taken to be approximately true, there seem to be two cases to consider. The first is one where the differences can be explained by the factors idealized or by experimental/measurement error. While my preference is still one of avoidance of approximate truth talk, it is this sort of case that is the least problematic for me.[30] The second sort of case is one where the "close" differences cannot be explained by appeal to the factors idealized nor by error theory. In these cases, there is something wrong with our lists that underlie the false equations. If one takes the theory to be the laws conjoined with lists as herein construed, then in such cases the theory is very likely false. There are unknown factors at work. My intuitions here are not at all

---

he does so in some instances in a manner almost identical to my talk about lists. He states that a proviso must provide ontic as opposed to epistemic completeness, and this means that it must provide "all the factors present in the given case which in fact affect the outcome to be predicted by the theoretical inference..." In other places his use of 'proviso' would not fit my sense of 'list'. See his "Provisos: A Problem Concerning the Inferential Function of Scientific Theories," *Erkenntnis*, xxviii (1988): 147–64.

[30] I have no trouble with approximate truth theories that try to make sense of such talk in terms of one equation being approximately true by comparison to the *true* equation. My problem consists in taking a false equation to be approximately true on the basis of the $|O - C|$ differences, as opposed to the *true* equation. It is the work of Ilkka Niiniluoto that made me see this distinction clearly—see his *Is Science Progressive?* (Boston: Reidel, 1984) and his *Truthlikeness* (Boston: Reidel, 1987).

clear. Here we have the theory likely to be false, the equations are false, the differences are not explained, and there are unknown factors at work. Is it at least a bit oversimplified or misleading to say under such conditions that the equation is approximately true?

Third, some will want to know whether I believe that there are mathematical laws, fundamental or otherwise, that are not based in some manner on false assumptions. Note that when I say "based on false assumptions," I mean either derived or constructed on the basis of such assumptions or requiring such assumptions for each application. Note also that from my perspective there can be *true fundamental theories* all of whose mathematical laws are based upon false assumptions. With this much said, it is certainly possible to have mathematical laws not based in some manner on false assumptions whose $|O - C|$ differences are explainable by appeal to measurement or experimental error. The fact that I have not found any such laws is not much of an argument against their possibility, given my limited experience and knowledge. My skepticism about the existence of such laws is buttressed by the writings of those like Cartwright, Laymon, Nowak, Giere, Ramsey, and others. For example, of even so fundamental a law as the Shrödinger equation Giere says that there is no real system modeled by the equation (since more details are needed including approximations). Because he does not believe that the Shrödinger equation describes a real system to start with, he, unlike Cartwright,[31] does not believe the equation "lies." With sufficient details it is used to define models.[32]

But for the sake of argument, what happens to my position if there are such laws in the form of equations that are both true and fundamental? There are two possible scenarios here. In the first case, the fundamental, empirical, nonvacuous, true equations with some suitable premises logically imply all of the vacuous, empirical mathematical laws (in my sense of 'law').[33] This is surely just plain false. But even if it were true, the false equations in the vacuous laws will not be deducible from the true fundamental equations. Such a deduction will require false assumptions (my ideal conditions). This situation would be the same as one where we take the Newtonian equations as true and fundamental and then try to assess the false Atwood

---

[31] *How the Laws of Physics Lie.*

[32] *Explaining Science*, p. 90.

[33] It is interesting to note that in "Laws of Science, Theories, Measurement: Comments on Ernest Nagel's *The Structure of Science*," Nowak criticizes Nagel on this very point in his discussion of *The Structure of Science* (New York: Harcourt, Brace, and World, 1961).

Machine equation. We would not escape the need to determine relevance of ideal conditions; we would still be stuck with assessing differences, and so on. In other words, we cannot deduce the vacuous laws in my sense from the true fundamental equations without false idealizing assumptions. The same problems remain for the (now) deduced vacuous laws as for those not based on fundamental true equations. These are the problems of "gap" size, relevance of the assumed conditions, best fitting function given the unidealized variables, complete lists, and so on. Of course, we are also stuck with these problems if some vacuous laws are not implied by nonvacuous fundamental laws. The acceptance/rejection problems for vacuous mathematical laws discussed in this article remain whether there are true fundamental equations or not.[34]

From my perspective, if one is to find nonvacuous mathematical laws it is more likely that they will not be fundamental. Instead, they will be the end result of Nowak's method of concretization. They will be laws where all ideal conditions have been dropped and corrected for such that all factors on the list are captured by the equation.

Finally, some further remarks comparing the approach of this paper with Giere's approach. As we have seen, Giere wants his model to be similar to the real system in certain respects. So far so good. But Giere does not see the need to specify the relevant respects in which the model is not similar to the real system. As such, he does not see that the ideal conditions fulfill this role of specifying "allegedly" relevant dissimilarities. From my perspective the ideal conditions have the role of marking the relevant respects in which the real system and the model are not similar. This of course raises the issue of relevance of the ideal conditions, and my treatment of this issue provides a general principle for saying that the expected differences are close enough. Giere seems not to address this issue of relevance and does not offer a *general* principle or account for when the differences or degrees are close enough. To repeat one more time, the ideal conditions are relevant only if the factors they idealize exist and explain the differences in whole or in part.

JAMES DERDEN

Humboldt State University

---

[34] I do not speculate on the possibility of having a true fundamental equation that implies a host of "bad" vacuous laws and how one would then assess the true equation in these circumstances.

BOOK REVIEWS

Knowledge and Its Limits. TIMOTHY WILLIAMSON. New York: Oxford University Press, 2000. xi + 340 p. Cloth $39.95, paper $19.95.

Timothy Williamson is perhaps best known for his views concerning the epistemic nature of vagueness. In the present book, knowledge itself comes under his microscope. His book articulates and defends a conception of the nature of knowledge—or at least, knowing *that*; it has little to say on knowing *how*, knowing *who*, and so forth[1]—and applies the result to a variety of related issues. It refashions and melds together a dozen or so papers that Williamson has written on the topic in the last ten years, to produce a very distinctive vision of the nature of knowledge. The book displays notable originality, and is carefully and cleverly argued. The resources of formal logic and probability theory are brought to bear where appropriate, more technical material of this kind being relegated to five appendices.

Traditional accounts factorize knowledge into two components, a subjective one, belief, plus an objective one, the appropriate relation to the world. This book rejects such an analysis. Knowledge is a *sui generis* mental state. Chapter 1 explains this conception of knowledge, and attacks various arguments to the effect that knowledge is not a purely mental state—as well as some claims concerning the nature of first-person authority on which such arguments depend. Chapter 2 disposes of some further arguments to the same end. Whether a mental state is one of knowledge certainly does depend on the cooperation of the world; however, the thought that the contents of our mental states depend on the external world is now a familiar one to philosophers, and, the chapter argues, the dependence of knowledge on the external world requires an externalism no more objectionable than that. Chapter 3 then goes on to the attack, and provides an ingenious argument against any analysis of knowledge into internal and external components.

Having defended his conception of knowledge from infanticide, the next several chapters go on to elaborate its properties. Chapter 4 argues that, quite generally, one's mental states are such that one cannot always know when one is in them. The argument is a sorites-

---

[1] But see Jason Stanley and Williamson, "Knowing How," this JOURNAL, XCVIII, 8 (August 2001): pp 411–44.

0022-362X/03/0005/268–71

like argument, and depends on the margin of error principle, of which more anon. The last couple of pages of the chapter draw, almost nonchalantly, a simple corollary which, if right, lays waste Dummettian antirealism. Chapter 5 draws another corollary of the argument, the rejection of the KK principle (if you know something, then you know that you know it), and discusses various connections between knowledge, reliability and the margin of error principle. Chapter 6 applies the results of these cogitations to the surprise exam paradox, arguing that the students in question start with the knowledge that there will be a surprise exam, but lose it as the days go on—and not because they forget things. Chapter 7 attacks "sensitive" accounts of knowledge—that is, those according to which, if you know p then the following conditional is true: if p had not been true, you would not have believed that p. It rejects this conditional and its variations.

Chapter 8 takes skepticism in its sights. Consider two people—one is you sitting in a chair; the other is a brain in a vat receiving identical sensory stimulation. Both of you believe yourselves to be sitting in a chair. According to the account of knowledge just developed, because you have the appropriate relation to the external world, you know that you are sitting in a chair, whilst your envatted doppelgänger does not. Skepticism solved. Since knowledge must be based on evidence, it follows that the two parties do not have the same evidence. It follows that evidence is not to be identified with that with which one is presented phenomenologically. What, then, is evidence? Chapter 9 argues that something is evidence if and only if it is known.

These conclusions challenge certain standard ideas concerning probability and evidence, and specifically the thought that we update our beliefs by conditionalization. If this were correct, knowledge could not be lost (as we have seen it can be). Chapter 10 provides a different account of conditionalization: the novel probability of p is the old probability of p, conditional upon the conjunction of the new evidence and *what remains of* the old evidence (220). (This is one place where Williamson is less than clear. The phrase 'what remains of' is omitted, making the subsequent discussion hard to follow. The situation is not helped by a typo: 'e_w' on l. 6 should be 'e_n'.) Assuming that one should assert that for which one has evidence (and only that), it follows from the preceding considerations, that one should assert what one knows (and only that). This conclusion is defended in chapter 11.

Chapter 12, which has only a loose connection with what has gone before, defends the claim that there are unknowable truths, by articulating and defending the well-known Fitch argument to the effect that if one assumes that everything true is knowable, it follows that everything true is known.

Williamson's views are novel, provocative, and raise many interesting issues. There is no space here to take many of them up. In what follows, I will take up just one—though an important one. Perhaps the central feature of Williamson's view of knowledge is its strong externalism. You and the brain in the vat have the same phenomenological experiences, but you know that you are sitting in a chair whilst the brain does not. I want to bring this thought to bear on the margin of error principle, which, as stated by Williamson,[2] is as follows:

> '*A*' is true in all cases similar to cases in which 'It is known that *A*' is true (*ibid*, p. 227).

This principle is a central plank of Williamson's view concerning vagueness: it allows him to explain why we cannot know exactly where the sharp cut-off point is in a soritical progression, even though there is one. Suppose that a vague sentence, *A*, changes from true to false between two adjacent points, *i* and *j*. If one knew that *A* were true at *i*, by the margin of error principle, it would have to be true at *j*, which it is not.

The natural rationale for the margin of error principle depends on the fact that knowledge supervenes on evidence. If two situations present the same evidence, for example, by being phenomenologically indistinguishable, then I cannot know that *A* holds in one but not in the other. In other words, if *A* holds in one but not the other, then any judgment I make about it in these situations must be a guess. Now, in response to the skeptic, Williamson has argued that there is more to evidence than is phenomenologically present. Things can appear the same to two people phenomenologically; yet one can know something and the other not. The one phenomenology counts as evidence precisely because it is knowledge. In the same way, someone can have the evidence that *A* holds at *i*, but not have the evidence that it holds at *j* precisely because *A* is knowledge. Once we are strong externalists about the conditions for knowledge, the situations at *i* and *j* are *not* similar.

Williamson often defends the margin of error principle by appealing to reliability. For example, he says:

> The intuitive idea is that if one believes outright to some degree that condition *C* obtains, when in fact it does, and at a very slightly later time one believes outright on a very similar basis to a very slightly lower

---

[2] *Vagueness* (New York: Routledge, 1994).

> degree that *C* obtains, when in fact it does not, then one's earlier belief is not reliable enough to constitute knowledge. The earlier case is sufficiently similar to the later case (101).

Now, as we have seen, the situations at *i* and *j* may not be similar in the relevant sense. One might still argue, however, that the first belief is not knowledge since it is not reliable, as demonstrated, for example, by the fact that a person will not make the same judgment on different occasions. But it seems to me that there may well be people who do make the same judgment about some vague situation on different occasions—I have, to my own surprise, found myself doing this sometimes. And even if there are not such people, there could be. (Recall that the epistemicist about vagueness wants to argue not just that the cut-off point is unknown, but that it is unknow*able*.) Of course, different people may also make different judgments about the same situation. But that is okay too. The one whose judgment is right knows; the other does not. Even if there is variation in the judgments of individuals, or of the same individual on different occasions, that is still okay. It may just show that vague words have a certain kind of context-dependence (not an implausible view). The person, relative to the context, may make the right judgment, and know that they have done so, even if the judgments are different in different contexts. Differences in judgments, therefore, even if they occur, are quite compatible with reliability. It would seem, then, that the epistemicist's argument as to why we cannot know the location of a cut-off point is in trouble.

As I pointed out above, the margin of error principle is also employed by Williamson to argue for various of his claims about knowledge. It would also seem, therefore, that his view about the nature of knowledge also threatens to undercut some of those arguments. That does not, of course, show that the view is wrong. And I think that Williamson's view has sufficient attractiveness and robustness to survive a problem of this kind. Williamson has shown us what it is like to be a serious and thoroughgoing externalist about knowledge. Before reading the book I would not have thought such a view a serious possibility. Now I do.

GRAHAM PRIEST

University of Melbourne

IN MEMORIAM: JAMES J. WALSH

The Editors and Trustees of the *Journal of Philosophy* note with sadness the passing of their friend and colleague, James J. Walsh, who died, abruptly, on February 24, 2003, of heart failure. Except for a period as Rhodes Scholar at Oxford University, his entire academic career, from 1954 until 1990, was in the Philosophy Department at Columbia, in which he served as Chairman, and as Director of Graduate Studies. He became an editor of the *Journal* in 1964. He brought to these offices the dedication and sense of duty that had made him a staff sergeant in the 34th Infantry Division, in which he was a radio operator in the front lines of Northern Italy in World War II. He must have been an exemplary soldier, for his constant impulse was to sacrifice his own interests to those of the organizations he served. Jim was a medievalist, and lived on terms of appreciative familiarity with his philosophical heroes—Duns Scotus, William of Ockham, and John Buridan—and he delighted in the ingenuity with which they were able to negotiate logical difficulties without the powerful tools that modern logic would have given them. He retired from teaching in 1990, intending to pursue his investigations into fourteenth-century thought, but also to live closer to nature. Jim came from the Pacific Northwest, where he had attended Reed College, and Manhattan held no charms for him. It was one of the paradoxes of his life that he sought to live as an outdoorsman—he was an avid skier and fisherman—while teaching in one of the most urban of Eastern universities. He spent a lot of his free time painting and writing poetry, but when his wife, Carole, was confined to a nursing home, he devoted himself to her care with the same intensity of love and what he lightly called *Pflicht*, that defined his relationship to everyone who depended upon him—his colleagues, his students, his cohorts in the infantry, and his children. He died in the snow, the element in which he was happiest. He was 78.

Arthur C. Danto
Bernard Berofsky
Isaac Levi
Charles D. Parsons
for the *Journal of Philosophy*

NEW BOOKS: TRANSLATIONS

ADORNO, THEODOR W.: *Essays on Music.* RICHARD LEPPERT, ed., SUSAN H. GILLESPIE and others, trans. Berkeley: California UP, 2002. xvii, 743 p. Cloth $75.00, paper $34.95.

ANSELM: *Three Philosophical Dialogues: On Truth, On Freedom of Choice, On the Fall of the Devil.* THOMAS WILLIAMS, trans. Indianapolis: Hackett, 2002. xiv, 110 p. Cloth $34.95, paper $9.95.

AQUINAS, THOMAS: *The Treatise on Human Nature: Summa Theologiae 1A 75-89.* ROBERT PASNAU, trans. Indianapolis: Hackett, 2002. xxi, 434 p. Cloth $39.95, paper $14.95; *Political Writings.* R.W. DYSON, ed. and trans. New York: Cambridge, 2002. xli, 318 p. Cloth $60.00, paper $21.00.

ARISTOTLE: *Nicomachean Ethics.* CHRISTOPHER ROWE, trans., SARAH BROADIE, ed. New York: Oxford, 2002. x, 468 p. Paper $19.95; *On Poetics.* SETH BENARDETE and MICHAEL DAVIS, trans. South Bend, IN: St. Augustine's, 2002. xxx, 105 p. Cloth $22.00, paper $10.00.

BADIOU, ALAIN: *Ethics: An Essay on the Understanding of Evil.* PETER HALLWARD, trans. New York: Verso, 2001. lviii, 166 p. Paper $16.00.

BALIBAR, ÉTIENNE: *Politics and the Other Side.* CHRISTINE JONES, JAMES SWENSON, and CHRIS TURNER, trans. New York: Verso, 2002. xv, 176 p. Paper $21.00.

BENTHIEN, CLAUDIA: *Skin: On the Cultural Border Between Self and the World.* THOMAS DUNLAP, trans. New York: Columbia, 2002. x, 290 p. Cloth $29.50.

BLANCHOT, MAURICE: *The Book to Come.* CHARLOTTE MANDELL, trans. Stanford: University Press, 2003. xii, 267 p. Cloth $55.00, paper $21.95.

BURIDAN, JOHN: *Summulae de Dialectica.* GYULA KLIMA, trans. New Haven: Yale, 2001. lxii, 1032 p. Cloth $120.00.

CASTORIADIS, CORNELIUS: *On Plato's Statesman.* DAVID AMES CURTIS, ed. and trans. Stanford: University Press, 2002. xxviii, 227 p. Cloth $55.00, paper $21.95.

CAVALIERI, PAOLA: *The Animal Question: Why Nonhuman Animals Deserve Human Rights.* CATHERINE WOLLARD, trans. New York: Oxford, 2001. viii, 184 p. Cloth $22.00.

CAVARO, ADRIANA: *Stately Bodies: Literature, Philosophy, and the Question of Gender.* ROBERT DE LUCCA and DEANNA SHEMEK, trans. Ann Arbor. Michigan UP, 2002. xiv, 220 p. Cloth $60.00, paper $24.95.

CHALIER, CATHERINE: *What Ought I to Do? Morality in Kant and Levinas.* JANE MARIE TODD, trans. Ithaca, NY: Cornell, 2002 197 p. Paper $17.95.

CHRÉTIEN, JEAN-LOUIS: *The Unforgettable and the Unhoped For.* JEFFREY BLOECHL, trans. New York: Fordham, 2002. xxii, 135 p. Cloth $35.00, paper $20.00.

CHUNG, BONGKIL, trans.: *The Scriptures of Won Buddhism: A Translation of the Wŏnbulgyo Kyojŏn with Introduction.* Honolulu: Hawaii UP, 2003. xviii, 413 p. Cloth $40.00.

COICAUD, JEAN-MARC: *Legitimacy and Politics: A Contribution to the Study of Political Right and Political Responsibility.* DAVID AMES CURTIS, trans. New York: Cambridge, 2002. xxvii, 259 p. Cloth $70.00, paper $25.00.

DERRIDA, JACQUES: *Ethics, Institutions, and the Right to Philosophy.* PETER PERICLES TRIFONAS, ed. and trans. Lanham, MD: Rowman and Littlefield, 2002. xi, 111 p. Cloth $60.00, paper $16.95; *Negotiations: Interventions and Interviews, 1997-2001.* ELIZABETH ROTTENBERG, ed. and trans. Stanford: University Press, 2002. viii, 404 p. Cloth $65.00, paper $24.95.

DESCOMBES, VINCENT: *The Mind's Provisions: A Critique of Cognitivism.* STEVEN ADAM SCHWARTZ, trans. Princeton: University Press, 2001. xxii, 284 p. Cloth $35.00.

DROIT, ROGER-POL: *The Cult of Nothingness: The Philosophers and the Buddha.* DAVID STREIGHT and PAMELA VOHNSON, trans. Chapel Hill: North Carolina UP, 2003. xii, 263 p. Cloth $59.95, paper $24.95.

DÜNKELSBÜHLER, ULRIKE OUDÉE: *Reframing the Frame of Reason: "Trans-lation" in and Beyond Kant and Derrida.* MAX STATKIEWICZ, trans. Amherst, NY: Humanity, 2002. 277 p. Cloth $65.00.

FERRY, LUC: *Man Made God: The Meaning of Life.* DAVID PELLAUER, trans. Chicago: University Press, 2002. 172 p. Paper $14.00.

FLAHAULT, FRANÇOIS: *Malice.* LIZ HERON, trans. New York: Verso, 2003. x, 193 p. Paper $20.00.

FUBINI, RICCARDO: *Humanism and Secularization: From Petrarch to Valla.* MARTHA KING, trans. Durham, NC: Duke, 2003. viii, 306 p. Cloth $64.95.

GADAMER, HANS-GEORG: *The Beginning of Knowledge.* ROD COLTMAN, trans. New York: Continuum, 2002. 148 p. Cloth $24.95.

GENTILE, GIOVANNI: *Origins and Doctrine of Fascism with Selections from Other Works.* JAMES GREGOR, trans. New Brunswick, NJ: Transaction, 2002. xiv, 107 p. Cloth $39.95.

GEORGE, DAVID B., and JOHN R. FORTIN, trans. and eds: *The Boethian Commentaries of Clarembald of Arras.* Notre Dame: University Press, 2002. xxxi, 153 p. Cloth $40.00.

HADOT, PIERRE: *What Is Ancient Philosophy?* MICHAEL CHASE, trans. Cambridge: Harvard, 2002. xii, 362 p. Cloth $29.95.

HAN, BÉATRICE: *Foucault's Critical Project: Between the Transcendental and the Historical.* Stanford: University Press, 2002. xiii, 241 p. Cloth $55.00, paper $24.95.

HEIDEGGER, MARTIN: *The Essence of Human Freedom: An Introduction to Philosophy.* TED SADLER, trans. New York: Continuum, 2002. xiv, 216 p. Cloth $99.95, paper $29.95; *The Essence of Truth: On Plato's Cave Allegory and Theaetetus.* TED SADLER, trans. New York: Continuum, 2002. xii, 252 p. Cloth $99.95, paper $29.95; *Off the Beaten Track.* JULIAN YOUNG and KENNETH HAYNES, eds. and trans. New York: Cambridge, 2002. x, 291 p. Cloth $60.00, paper $22.00.

HENRY, MICHAEL: *I Am the Truth: Toward a Philosophy of Christianity.* SUSAN EMANUEL, trans. Stanford: University Press, 2003. ix, 284 p. Cloth $60.00, paper $21.95.

HERDER, JOHANN GOTTFRIED: *Sculpture: Some Observations on Shape and Form from Pygmalion's Creative Dream.* Chicago: University Press, 2002. ix, 141 p. Paper $18.00.

OUTSTANDING SCHOLARSHIP FROM CAMBRIDGE

The Improvisation of Musical Dialogue
A Phenomenology of Music
Bruce Ellis Benson
0-521-81093-0 Hardback $53.00
0-521-00932-4 Paperback $19.00

Symbolic Representation in Kant's Practical Philosophy
Heiner Bielefeldt
0-521-81813-3 Hardback $60.00

The Cambridge Companion to the Scottish Enlightenment
Edited by Alexander Broadie
Cambridge Companions to Philosophy
0-521-80273-3 Hardback $60.00*
0-521-00323-7 Paperback $22.00*

Time, Reality and Experience
Edited by Craig Callender
Royal Institute of Philosophy Supplements 50
0-521-52967-0 Paperback $20.00

Freedom and Anthropology in Kant's Moral Philosophy
Patrick R. Frierson
0-521-82400-1 Hardback $55.00*

The Cambridge History of Seventeenth-Century Philosophy
Volumes 1 and 2
Edited by Daniel Garber and Michael Ayers
0-521-53180-2
2-Volume Paperback Set $75.00

The Cambridge Companion to Bertrand Russell
Edited by Nicholas Griffin
Cambridge Companions to Philosophy
0-521-63178-5 Hardback $65.00*
0-521-63634-5 Paperback $23.00*

Michel de Montaigne
Accidental Philosopher
Ann Hartle
0-521-82168-1 Hardback $60.00

Mind, Reason and Imagination
Selected Essays in Philosophy of Mind and Language
Jane Heal
Cambridge Studies in Philosophy
0-521-81697-1 Hardback $75.00
0-521-01716-5 Paperback $28.00

Hellenistic and Early Modern Philosophy
Edited by John Miller and Brad Inwood
0-521-82385-4 Hardback $60.00

States, Nations and Borders
The Ethics of Making Boundaries
Edited by Margaret Moore and Allen Buchanan
0-521-81971-7 Hardback $65.00
0-521-52575-6 Paperback $24.00

Descartes's *Cogito*
Saved from the Great Shipwreck
Husain Sarkar
0-521-82166-5 Hardback $65.00

**Prices subject to change*

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

the review of

metaphysics

a philosophical quarterly

MARCH 2003 | **VOL. LVI, No. 3** | **ISSUE No. 223** | **$15.00**

articles

GORDON BARNES *The Paradoxes of Hylomorphism*

ELISABETH UFFENHIEMER-LIPPENS *Rationalized Passion and Passionate Rationality: Thomas Aquinas on the Relationship between Reason and the Passions*

ANN MARGARET BAXLEY *Does Kantian Virtue Amount to More than Continence?*

HERMAN DE DIJN *Hume's Nonreductionist Philosophical Anthropology*

DAVID JANSSENS *The Problem of Enlightenment: Strauss, Jacobi, and the Pantheism Controversy*

ELIZABETH C. SHAW AND STAFF *Summaries and Comments*

book reviews *philosophical abstracts* *announcements*

Individual Subscriptions $35.00 Institutional Subscriptions $55.00 Student/Retired Subscriptions $20.00

VISA AND MASTERCARD ACCEPTED

Jude P. Dougherty, Editor
The Catholic University of America, Washington D.C. 20064
email: mail@reviewofmetaphysics.org

PB Prometheus Books

SCIENCE AND RELIGION

Are They Compatible?

Edited by Paul Kurtz

with the assistance of Barry Karr and Ranjit Sandhu

"... science and religion have been seeking common ground through ongoing dialogue.... this volume provide[s] a dimension to the conversation that has seldom been heard.... present[s] important and provocative voices too often drowned out."

—*Publishers Weekly*

368 pp • ISBN 1-59102-064-6 • PB $20

HAPPINESS IN A GODLESS WORLD

George H. Smith

If you don't believe in God or an afterlife, how can you possibly lead a happy life? Secular humanist writer George H. Smith tackles this question in this stimulating overview of major secular theories of happiness.

250 pp • ISBN 1-59102-038-7 • PB $19

HAS SCIENCE FOUND GOD?

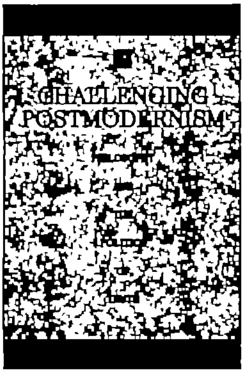
The Latest Results in the Search for Purpose in the Universe

Victor Stenger

In the past few years a number of scientists have claimed that there is credible scientific evidence for the existence of God. Is this true? Are scientists close to solving the greatest of all mysteries? Physicist Victor J. Stenger delves into this fascinating question from a skeptical point of view in this lucid and engrossing presentation of the key scientific facts.

295 pp • ISBN 1-59102-018-2 • HC $30

HB Humanity Books



CHALLENGING POSTMODERNISM

Philosophy and the Politics of Truth

David Detmer

While sympathetic to Continental philosophy, Detmer nonetheless defends the idea of objective truth and argues that doing so is a matter of considerable political importance.

320 pp • ISBN 1-59102-101-4 • PB $35

TRANSLATING HEIDEGGER

Miles Groth

Groth examines both the history of the first English translations of Heidegger's works and Heidegger's philosophy of translation, revealing that there is a coherent philosophy of translation in Heidegger's texts.

230 pp • ISBN 1-59102-100-6 • HC $60

For exam copies contact editorial@prometheusbooks.com

Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com

Philosophy

NEW FROM THOEMMES PRESS

AN ENTIRE BODY OF PHILOSOPHY, ACCORDING TO THE PRINCIPLES OF THE FAMOUS RENATE DES CARTES
Antoine le Grand
804 PAGES, 2 VOLUMES 1-84371-052-8
CLOTH $375.00

THE COLLECTED WRITINGS OF ADDISON W. MOORE
Addison W. Moore
Edited and with an Introduction by John R. Shook, with additional Introductions by J. Brent Crouch and Scott Pratt
850 PAGES, 3 VOLUMES 1-84371-002-1
CLOTH $335.00

***THE DANGER OF WORDS* AND WRITINGS ON WITTGENSTEIN**
1973 Edition
Maurice O'Connor Drury
Edited by David Berman, Michael Fitzgerald, and John Hayes
334 PAGES 1-84371-045-5
CLOTH $95.00

MISCELLANEOUS WRITINGS, SPEECHES AND LETTERS
T.H. Green
Edited by Peter Nicholson
508 PAGES 1-84371-014-5
CLOTH $125.00

WINCKELMANN UND SEINE ZEITGENOSSEN
Winckelmann and his Contemporaries
1898 Edition
Carl Justi
1254 PAGES, 3 VOLUMES 1-84371-005-6
CLOTH $335.00

THE FIRST MATE'S LOG
1940 Edition
R.G. Collingwood
With an Introduction by Peter Nicholson
263 PAGES 1-85506-328-X
CLOTH $95.00

www.thoemmes.com

DISTRIBUTED IN NORTH AMERICA BY THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu

Original.

New in paperback

Freedom and Its Betrayal

Six Enemies of Human Liberty

Isaiah Berlin

Edited by Henry Hardy

"When reading Isaiah Berlin, we breathe an altogether different air, and not simply because he was a superior writer. With him we know we are inside the psychological and historical clockwork that turns the hands of modern political life."
—Mark Lilla, *New York Review of Books*

Paper $17.95 ISBN 0-691-11499-4

New in paperback

Heidegger's Children

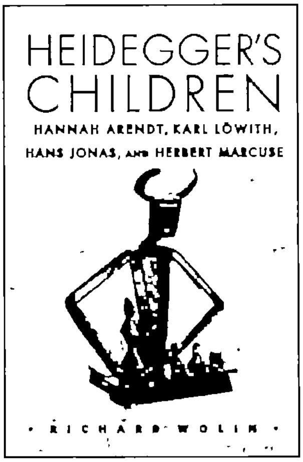
Hannah Arendt, Karl Löwith, Hans Jonas, and Herbert Marcuse

Richard Wolin

"*Heidegger's Children* presents an articulate and convincing account of the moral and political weaknesses of Heidegger's philosophy as well as highly informative studies of his influence on four of the twentieth century's most important philosophers."
—Brian J. Fox, *Review of Metaphysics*

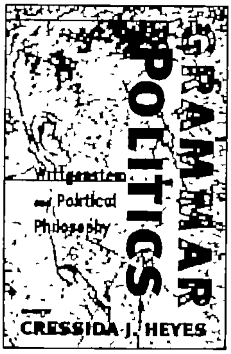
"A most thought-provoking and illuminating look at Heidegger's legacy."
—*Choice*

Paper $17.95 ISBN 0-691-11479-X



Princeton University Press

800-777-4726 • WWW.PUP.PRINCETON.EDU



The Grammar of Politics

Wittgenstein and Political Philosophy

EDITED BY CRESSIDA J. HEYES

The Grammar of Politics demonstrates the variety of ways political philosophers understand Wittgenstein's importance to their discipline and apply Wittgensteinian methods to their own projects. $45.00 cloth, $19.95 paper

Jean-Jacques Rousseau

A Friend of Virtue

JOSEPH R. REISERT

"Joseph Reisert's book is a compact, interesting analysis of Rousseau and of the place that virtue plays in his diagnosis of and prescription for modern society."—Tracy B. Strong, University of California, San Diego $42.50

Beautiful City

The Dialectical Character of Plato's Republic

DAVID ROOCHNIK

"While Plato's *Republic* is surely one of the richest books in the history of philosophy, it is also one of the most written about, and therefore presents a formidable challenge: how to say anything new? Only a genuinely new 'take' on the dialogue will allow for a new vein of its richness to be tapped; David Roochnik has succeeded admirably in doing just that."
—Drew Hyland, Trinity College. $35.00

Avicenna's Metaphysics in Context

ROBERT WISNOVSKY

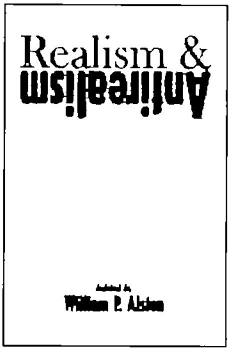
"Robert Wisnovsky's judgments in [this book] are based on a keen critical sense and sound knowledge of all the relevant texts, Greek and Arabic alike. The combination of historical and philological accuracy and philosophical insight make this book relatively rare among its kind in Arabic philosophy."—Dimitri Gutas, Yale University. $65.00

Ways of Being

Potentiality and Actuality in Aristotle's Metaphysics

CHARLOTTE WITT

"This unique book discusses Aristotle's theory of value and connects it in a very illuminating way to contemporary feminist critical readings of Aristotle's metaphysics."—Theodore Scaltsas, University of Edinburgh. $35.00



Realism and Antirealism

EDITED BY WILLIAM P. ALSTON

"*Realism and Antirealism* looks at the issues from a number of different angles, and arrives at a variety of fascinating conclusions. The book . . . makes a genuine contribution to the discussion of realism and antirealism."
—Nicholas Wolterstorff, Yale University
$45.00 cloth, $22.50 paper

Cornell University Press

At bookstores, or call (607) 277-2211 • www.cornellpress.cornell.edu

On Escape
De l'évasion

EMMANUEL LEVINAS
Translated by BETTINA BERGO
Introduced and Annotated by
JACQUES ROLLAND

First published in 1935, *On Escape* represents Emmanuel Levinas's first attempt to break with the ontological obsession of the Western tradition. In it, Levinas not only affirms the necessity of an escape from being, but also gives a meaning and a direction to it.

$14 95 paper $35 00 cloth

Deconstruction and the Remainders of Phenomenology
Sartre, Derrida, Foucault, Baudrillard

TILOTTAMA RAJAN

"A superbly strong work that will generate much discussion. There are no comparable books that have undertaken such a rereading of the development of French theory."
—David F. Ferris,
University of Colorado at Boulder

$25 95 paper $65 00 cloth

The Politics of Aesthetics
Nationalism, Gender, Romanticism

MARC REDFIELD

"This is a remarkable book of penetrating intelligence. Through a series of deft readings, Redfield shows us how the aesthetic is always entangled in lines of forces beyond art proper."
—Ian Balfour, York University

$55 00 cloth

Husserl's Phenomenology

DAN ZAHAVI

Drawing upon both Husserl's published works and posthumous material, *Husserl's Phenomenology* incorporates the results of the most recent Husserl research. It can consequently serve as a concise and updated introduction to his thinking.

$15 95 paper $45 00 cloth

Can One Live after Auschwitz?
A Philosophical Reader

THEODOR W. ADORNO, Edited by
ROLF TIEDEMANN, Translated by
RODNEY LIVINGSTONE and Others

"Despite his conviction that no philosophy could presume to approach an event like Auschwitz, this collection of Adorno's essays and aphorisms attests to his extraordinary effort to regard human suffering as the precondition of thought and as the undoing of all claims to totality."
—Anson Rabinbach,
Princeton University

$29 95 paper $75.00 cloth

The Idea of Form
Rethinking Kant's Aesthetics

RODOLPHE GASCHÉ

"The book does a wonderful job of unfolding and disentangling Kant's dense and very complex arguments. In addition to philosophers, it will be of interest to those working in literary criticism, art theory, and political theory."
—Dennis Schmidt,
Villanova University

$24 95 paper $55 00 cloth

Democracy in Question
Democratic Openness in a Time of Political Closure

ALAN KEENAN

This book explores the theoretical paradoxes and practical dilemmas that flow from the still radical idea that in a democracy it is the people who rule, and argues that accepting the open and uncertain character of democratic politics can lead to more sustainable and widespread forms of democratic engagement.

$19 95 paper $49 50 cloth

Stanford
University Press
800.621.2736 | www.sup.org

Hans-Georg Gadamer

A Biography

JEAN GRONDIN

Translated by Joel Weinsheimer

This book—the **first full-length biography of Gadamer in English**—"is a welcome and gracious gift to those of us eager to understand as much as possible about one aspect of Gadamer's hermeneutic philosophy, that is, the man himself who was its occasion."
—Francis J. Ambrosio

Yale Studies in Hermeneutics 30 illus. $35.00

The Cabala of Pegasus

GIORDANO BRUNO

Translated and annotated by
Sidney L. Sondergard and Madison U. Sowell

"**An important addition to the understanding of Bruno** and of Renaissance Hermeticism generally."
—Thomas Willard $40.00

Zarathustra's Secret

The Interior Life of Friedrich Nietzsche

JOACHIM KOHLER

Translated by Ronald Taylor

"**This important and boldly unique book** supplements all those strictly philosophical studies of Friedrich Nietzsche that have excluded his sexuality."
—*Library Journal* 43 illus. $29.95

Problems of Idealism

Essays in Russian Social Philosophy

Translated, edited, and introduced by
RANDALL A. POOLE

Foreword by Caryl Emerson

"[This] contribution of first importance ... [presents] articles of such major figures as **Berdiaev, Bulgakov, Frank, Kistiakovskii, and Novgorodtsev.**"
—Richard Wortman

Russian Literature and Thought Series $55.00

New in paper

The Praise of Folly

Second Edition

DESIDERIUS ERASMUS

Translated, with an introduction and commentary by Clarence H. Miller, with an afterword by William H. Gass

"**Exciting and brilliant, this is likely to be the definitive translation** of *The Praise of Folly* into English."—Richard J. Schoeck (on the earlier edition)

$12.95 A Yale Nota Bene paperback

Way to Wisdom

An Introduction to Philosophy

Second Edition

KARL JASPERS

Translated by Ralph Manheim

With a new foreword by Richard M. Owsley

"**Original, sincere, cultivated, and stimulating.**"—*Philosophy*

$14.95 paperback

On Liberty

JOHN STUART MILL

With essays by David Bromwich, Jean Bethke Elshtain, Owen Fiss, George Kateb, Richard A. Posner, and Jeremy Waldron

This classic work is now available in **a new edition that includes essays by distinguished scholars.**

Rethinking the Western Tradition

$26.00 cloth; $10.95 paperback

Hermeneutics as Politics

STANLEY ROSEN

With a foreword by Robert B. Pippin

"**Perhaps the most original and philosophically important** critical account of hermeneutics—of its philosophical status and historical development—to appear since Gadamer's *Truth and Method*."
—*Choice* $20.00 paperback

Yale University Press

yalebooks.com

THE JOURNAL OF
PHILOSOPHY

২৭।৪।০৩

VOLUME C, NUMBER 6
JUNE 2003
SPECIAL ISSUE: ON BERNARD WILLIAMS

page
275 *Thickness and Theory* T. M. Scanlon
288 *Reasons Thin and Thick* Allan Gibbard
305 *Ethics and Ontology* Charles Taylor

BOOK REVIEWS
321 Donald Davidson: *Subjective, Intersubjective, Objective* Frank Jackson

326 NEW BOOKS: TRANSLATIONS
328 NEW BOOKS: PAPERBACK REPRINTS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2003 by the Journal of Philosophy, Inc ISSN 0022-362X

২৮৫ ২৯।৮।০

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 6, JUNE 2003

THICKNESS AND THEORY*

When I read Bernard Williams's[1] trenchant objections to moral theory and "the morality system," I am torn, and uncertain where I stand. A kind of existential crisis ensues. Am I a defender of moral theory? Should I respond defensively: "No, that's not what a moral theorist should think. Morality doesn't need to be like that!" Or should I take a more detached view: "Yes, moral theory and the morality system are mistaken. I'm glad I don't believe any of that!"

What I will do here is to go through some of the leading ideas in Williams's critique of moral theory, trying at the same time to link them together and to work out a stable set of reactions to them. I will begin with the contrast between *thick* and *thin* concepts, and continue from there to questions about the nature and viability of moral theory.

Williams famously says that the application of thick ethical concepts such as "coward, lie, brutality, gratitude and so forth" is both action-guiding and world-guided. "If a concept of this kind applies," he writes, "this often provides someone with a reason for action.... At the same time, their application is guided by the world. A concept of this sort may be rightly or wrongly applied, and people who have acquired it can agree that it applies or fails to apply to some new situation" (*ibid.*, pp. 140ff.).

This world-guidedness might be thought to be a matter of the greater empirical content of thick concepts as contrasted with thin

---

*[The papers in this special issue were delivered at a conference in honor of Bernard Williams, sponsored by this JOURNAL, and held at Columbia University on November 13–14, 1999. We thank all the participants and all who attended.]

[1] *Ethics and the Limits of Philosophy* (Cambridge: Harvard, 1985).

0022-362X/03/0006/275–87

ones like 'right' and 'ought'. On this understanding of the term, a concept is "world-guided" to the extent that its correct applicability to a situation is determined by empirical (physical and psychological) features of that situation. When the concept is understood in this way, however, the significance of the contrast between thick and thin concepts seems open to doubt. As Samuel Scheffler[2] pointed out in his review of Williams's book, ethical theories typically deal with a variety of moral concepts including well-being, consent, and various specific rights, such as rights to privacy, all of which seem to have considerable empirical content. So the contrast in empirical content between thick and thin concepts is at most one of degree.

Moreover, the term 'world-guided' suggests that what distinguishes thick concepts is a kind of objectivity, that judgments using these concepts can express knowledge because (in virtue of their thickness) they make claims about the world—claims that are made true by (hence their proper use "guided by") what the world is like. But there is a doubt about this claim of objectivity, which is raised by something Williams himself says. Responding to an imagined challenge from a noncognitivist, that thick ethical concepts can be divided into separate descriptive and evaluative components, Williams points out that there is no reason to believe that there will be a descriptive concept that will pick out just the same features of the world as a world-guided thick ethical concept. In order to trace the contours of the ethical concept's applicability we have to understand its evaluative point. But if this is true—if we must be guided by the evaluative perspective of a thick concept in order to apply it—this seems to undermine the claim that its application is guided by the world in a sense that, by itself, conveys objectivity. To claim objectivity for these concepts we would also need, it would seem, to claim it for the evaluative judgments that determine their empirical shape.

This suggests a different way of understanding the contrast between thick and thin ethical concepts, one that emphasizes not the degree of empirical content of thick concepts, but their degree of content *period*, especially the richness of the evaluative perspective that they involve. Here, it will be helpful to start from a characterization of thin concepts. The thinness of thin concepts such as 'right', 'wrong', 'duty', and 'obligation' lies first and foremost in the abstractness, hence relative emptiness, of the ethical ideas that they involve, by contrast with concepts such as 'coward', 'lie', 'betrayal', 'brutality',

---

[2] "Morality through Thick and Thin," *Philosophical Review*, XCVI (1987): 411–34, see especially pp. 417–18.

'honor', and 'gratitude', whose greater content reflects the distinctive character of particular social worlds. Differences in the degree of empirical content, and hence in "world-guidedness," are consequences of this deeper difference.

This interpretation of the contrast fits with something that Williams says in later work[3] about why judgments employing thick concepts might constitute knowledge, while those employing thin concepts could not. He says that there could be such a thing as expertise in the application of thick concepts, but no analogue of this in the case of thin ones. Expertise becomes relevant to knowledge through the following argument.

In order for a person to know that *p*, *p* must be true, he must believe that *p*, and he must be warranted in this belief. The importance of this third condition, Williams says, citing E. J. Craig, "has its roots in the fact that we need some detectable property to lead us to informants with reliable information."[4] He goes on to say that with respect to some kinds of questions, including empirical matters and theoretical questions in science and mathematics, people have different levels of expertise because they have different opportunities for becoming acquainted with the relevant information. There are two kinds of possible expertise here (corresponding to two kinds of warrant and hence two kinds of knowledge). The first is expertise grounded in training in relatively abstract theoretical matters, such as quantum mechanics. The second is expertise that consists in having the right kind of direct experience of the phenomena themselves.

There is, Williams argues, no analogue of theoretical expertise in ethics. He gives two sorts of arguments for this. The first appeals to what he takes to be the absurdity of someone's believing an ethical claim on the authority of his teacher's credentials as an expert, even though he cannot see it for himself. The second is the chilling idea of life and death decisions in a hospital, for example, being made by someone whose only qualification is a Ph.D. in moral philosophy. These observations have, initially, considerable force. But what is to be made of them?

Is it implausible to think that someone could have good reason to accept an ethical conclusion simply because some moral philosopher whom he believed to have thought long and hard about the subject was convinced that it was true? It would no doubt be unfair here for me to say, but is nonetheless true, that the fact that *Williams* is convinced that some ethical claim is mistaken gives me pause in continuing to hold it—at the very least gives me reason to spend a lot of time reexamining it—even if I cannot, at first, see why he thinks this. And I believe that this has to do not only with the fact that Williams is Williams, but also with the fact that he is an excellent moral philosopher, not a philosopher of physics, a physicist, a literary critic, nor an anthropologist.

---

[3] "Truth in Ethics," in Brad Hooker, ed., *Truth in Ethics* (Cambridge: Blackwell, 1996), pp 19–34, and "Who Needs Ethical Knowledge?" in Williams's *Making Sense of Humanity and Other Philosophical Papers* (New York: Cambridge, 1995), pp. 203–12.

[4] "Who Needs Ethical Knowledge?" p. 205.

Supposing, however, that this claim about the implausibility of accepting an ethical conclusion "on authority" is correct, might there not be some explanation for it other than the irrelevance of moral theory? The implausibility of this kind of authority in ethics may rest on an idea about the kind of truths that ethical truths are supposed to be. They are supposed to be action-guiding and to serve as the basis for justifications and demands for justification that we can offer one another. If ethical judgments are to have this function, then it must not only be possible for us in general to see what they require of us, but also to see why this should be required. This explains our reaction to the idea of a student's taking his professor's word for an ethical conclusion, but it does not rule out theory, at least theory of a kind that meets the relevant transparency requirement. So it is not yet clear, to me at least, why there could not be something like theoretical expertise in ethics. This may become more clear, however, when we turn to Williams's reasons for thinking that there could be ethical expertise of the second kind he mentions.

This is the analogue of expertise grounded in experience. An adviser, Williams says, is someone who "may see better than you do how things stand and will help you see them aright." The adviser's expertise need not be only nonethical knowledge. It may, he writes, "take the form of certain kinds of knowledge under ethical concepts—that a certain course of action would be cowardly, for instance, or would count as a betrayal, or would not really be kind, and contributions of this kind can offer the person who is being advised a genuine discovery. So there is, in a sense, some ethical knowledge, it seems: knowledge of truths under 'thick ethical concepts'." There are also "marks of reliability" of such advisers, but these are "not best characterized in terms of possessing information, but rather in terms of certain capacities, such as judgment, sensitivity, imagination, and so forth."[5]

This may at first seem unfair, on two grounds. First, in order for the fact that a certain course of action would count as a betrayal to amount to a discovery, and one with reason-giving force, seeing that this is so would have to involve some new and better understanding of what makes something a betrayal. What the expert deploys thus begins to sound like a piece of ethical theory, as I would understand it. Second, the contrast proposed may strike one as a put-up job. The "theorist" is portrayed as someone who has learned some moral philosophy (can talk about the difference between deontological and consequentialist theories, for example) but is (utterly) lacking in judgment, while the possessor of thick ethical concepts is supposed also to be endowed with good judgment. Could we not have made the opposite assumptions—pairing a wise theorist with a woodenly conventional deployer of thick concepts?

---

[5] The quoted passages are from "Who Needs Ethical Knowledge?" pp. 206 and 207

I believe that, in Williams's view, there could be a woodenly conventional and unimaginative user of thick concepts, and that a wise person might also have a thorough understanding of moral theory, but this person's wisdom could not arise from or consist in this grasp of moral theory. This is because in Williams's view there is not enough *there*, at the level at which moral theory seeks to operate, for someone to be expert about. Here, we are brought back to the second interpretation of the distinction between thick and thin concepts that I discussed above. On this interpretation, the fundamental problem with thin concepts is a lack of the kind of distinctive evaluative point that gives thick concepts their definite shapes. Their lack of empirical content (and of the kind of objectivity that such "world-guidedness" entails) is one consequence of this, and, we now see, their failure to ground a form of expertise is another.

What this contrast between thick concepts and thin ones calls attention to is, in the first instance, an alleged weakness of thin concepts. The corresponding virtue of thick concepts is merely that they do not have this weakness. Emptiness is a fault, but content is a virtue only if it is the right content. There are plenty of thick concepts one would not want to employ. Concepts that presuppose racist views are an obvious case in point, but there is no shortage of other examples. So the question is, what kind of reasons can be given for using one set of thick concepts rather than another, or for understanding these concepts one way rather than another?

Williams agrees that we need to be able to give such reasons. He writes: "Nothing that has been said should lead us to think that traditional practices are beyond criticism; practices that make distinctions between different groups of people may certainly demand justification, if we are not to be content with unreflective traditions which can provide paradigms of prejudice."[6] The question, then, is how to distinguish between the kind of reflective criticism that Williams endorses and the kind that leads to ethical theory of the kind he urges us to reject. The difference lies, I believe, in the ways in which the mechanisms of criticism are related to the moral views (Williams often calls them *dispositions*) that they are used to criticize or support.

Theory of the kind Williams objects to looks for "justificatory reasons" for our ethical views (*ibid.*, p. 112). These are not just reasons for applying our thick ethical concepts in one way rather than another—reasons for calling certain actions "cruel," for example—but reasons for using these concepts at all and for taking seriously the judgments made with them. Such justifications can take either of two forms. What I will call *moral justification* tries to show that the judgments we make using our particular ethical concepts are correct judgments about what is morally right and wrong. *Rational justification*, on the other hand, tries to show that the judgments we make using certain ethical concepts are ones that we have good reason to take seriously as guides to action. This kind of justification might be sought even after adequate moral justification has been given. Even if the judgments we make using certain concepts do express genuine moral requirements, we may still ask, "Why care about that?"

Whichever kind of justification is in question, ethical theory tries to provide it by looking for a justificatory standpoint that is independent of our ethical concepts. Williams believes that this quest is mistaken. He writes:

> Any such picture makes in some degree the Platonic assumption that the reflective agent as theorist can make himself independent from the life and character he is examining. The belief that you can look critically at all your dispositions from the outside, from the point of view of the universe, assumes you could understand your own and other people's dispositions from that point of view without tacitly taking for granted a picture of the world more locally familiar than any that would be available from there; but neither the psychology nor the history of ethical reflection gives much reason to believe that the theoretical reasonings of the cool hour can do without a sense of the moral shape of the world, of the kind given in the ordinary dispositions (*ibid.*, p. 110).

If the evaluative attitudes that we express when we "look critically at our dispositions" are themselves included among our dispositions, then it is very plausible to claim that we cannot "look critically at all our dispositions from the outside" (at least not all of them at the same time). We will always, as it were, be looking through some of these dispositions rather than at them. Consider first the case of rational justification. The search for this kind of justification involves "standing outside of" certain of our judgments (those we call moral) and asking what reason we have for being guided by them. But in answering this question, we will still be expressing our own evaluative attitudes. I will call this the *no-escape thesis.*

---

[6] *Ethics and the Limits of Philosophy*, p. 115.

Some would hold, of course, that such an answer can be grounded in something more fundamental and inescapable than a particular evaluative judgment. Kant, for example, believed that his categorical imperative expressed a condition of rational agency: that recognizing its authority was not just a matter of coming to accept one evaluative judgment among others, but rather a matter of understanding what is required in order to have any such attitudes at all.

Like Williams, I am not convinced by Kant's argument, and I do not think that any other attempt to ground moral requirements in preconditions of rationality is likely to succeed. The alternative, as I see it, is what I call the *substantive strategy* of trying to characterize morality in a way that makes clear why it, and the kind of life it makes possible, are things that we have reason to value. This strategy runs the risk of failing to account for the special force of morality—"the moral must"—because it seems to base the claims of morality on what is just one good among many others. I think that the best response to this problem is to call attention to relations between this good and others that make it sensible to give it priority over them, rather than to try to argue that unlike other goods it is a requirement of rationality itself.[7]

In accepting the no-escape thesis, in rejecting Kant's argument, and in adopting this substantive strategy, I may be coming closer to the position Williams recommends in the following passage.

> The main consequence that this discussion has for ethical argument is that reflective criticism should basically go in a direction opposite to that encouraged by ethical theory. Theory looks characteristically for considerations that are very general and have as little distinctive content as possible, because it is trying to systematize and because it wants to represent as many reasons as possible as applications of other reasons. But critical reflection should seek for as much shared understanding as it can find on any issue, and use any ethical material that, in the context of the reflective discussion, makes some sense and commands some

---

[7] I defend this substantive strategy and try to carry it out in chapter 4 of *What We Owe to Each Other* (Cambridge: Harvard, 1998)

> loyalty. Of course that will take some things for granted, but as serious reflection it must know it will do that.[8]

Important points of disagreement may, however, remain between Williams's position and mine. What I have called a substantive strategy consists in trying to characterize an ethical concept in a way that brings out our reasons for taking it seriously as a guide to action.[9] This approach can be applied either to thick concepts or thin ones, but in my own work I have concentrated on thin concepts, specifically the concepts of right and wrong. Williams takes this emphasis to be characteristic of ethical theory of the kind he believes is misguided.

This brings me to what I called above moral justification. Williams sees ethical theory of the kind he opposes as essentially reductive. This is because it takes the process of critical reflection on our moral concepts to be a matter of determining whether the judgments made using these concepts are true judgments about what we ought morally to do, this being something that can be specified by general principles that use only thin concepts such as 'right' and 'wrong'. Williams believes that this approach is headed in the wrong direction. He believes, as we have seen, that judgments made using thick ethical concepts can be true in a substantial sense. But the various languages of assessment employing thick concepts are in his view not "homogeneous."[10] That is, they are not various ways of stating truths about one thing, "what we ought, morally, to do." Since judgments using these concepts all give, or purport to give, reasons for action, there is a sense of 'right' and 'wrong' in which they all make claims about what it is right or wrong to do. But this is just the sense (not specifically moral) of 'right' and 'wrong' in which to call an action "wrong" is to say that there are strong reasons (of some kind or another) against it, and to call an action "right" is just to say that there are no such reasons and, perhaps, that there are good reasons (of some kind or other) in its favor.[11] Williams doubts, however, that there is a more substantial, specifically moral, sense of right and wrong such that people in different cultures, or people within the same culture who belong to groups using different thick concepts, are all using those concepts to make claims about what is right or wrong in this sense.

---

[8] *Ethics and the Limits of Philosophy*, pp. 116–17.

[9] This inquiry may, of course, lead to a negative conclusion: that the concept in question does not provide good reasons, or that it would do so only if modified in certain ways.

[10] "Truth in Ethics," p. 33.

[11] This is my candidate for the sense that Williams is referring to when he speaks of "a univocal sense of 'right' and 'wrong' that is closely related to practical reason" (*ibid.*, p. 32).

Here, it is helpful to distinguish two questions. The first is why one would think that there are substantial notions of moral right and wrong that we have reason to take seriously. The second is whether these are the *only* ethical concepts that are, ultimately, to be taken seriously: whether we or others have good reason to take other ethical concepts as guides to action only if they are reducible to these concepts, or justified in terms of them.

In response to the first question, I believe that there is such a sense of moral right and wrong because it seems to me that an important part of our ethical thinking is unified in two respects. It seems to be unified in method—there is such a thing as the way that moral conclusions can and should be arrived at—and to be unified in its authority and in the motivation that it triggers: when one concludes that an action is wrong, this entails the recognition of a particular kind of reason not to do it. Moral theorists thus are driven to give a general characterization of the idea of moral rightness and wrongness (that is, of that part of our ethical thinking which has this unity) in order both to characterize the relevant method and to explain its rational authority.

An ethical theory that offers an account of this method may explain some thick ethical concepts in terms of thin concepts such as right and wrong. I believe, for example, that the obligation to keep a promise can be explained in terms of general principles arising from the interests that others have in being able to rely on expectations about what we are going to do.[12] These principles are more general than the concept of a promise: they apply to some situations in which expectations have been created without promises having been given as well as to ones in which promises have been made. I believe that the structure of the concept of a promise as we understand it, in particular the conditions under which a promise is binding, are well understood by asking what principles of this kind it would or would not be reasonable to reject. The point of view from which we answer this question is less "local" than the concept of a promise, if being "local" means presupposing a particular thick concept.[13] But it does presuppose certain features of our world that make the concept of a promise important for us, such as the fact that we often need to rely on assumptions about what others are going to do, and have reason to want to provide others with reason to rely on assumptions about our behavior. It thus may rely on what Williams calls "a sense of the moral shape of the world." If so, then what this example shows is that such a sense is not to be identified with the use of particular thick concepts.

---

[12] I offer such an explanation in chapter 7 of *What We Owe to Each Other*

[13] I have in mind here Williams's remark, quoted above, from p. 110 of *Ethics and the Limits of Philosophy*.

In a case of this kind, a moral theory explains a thick concept in terms of a thin one, so what it offers is in this respect a reductive account. But the result of coming to accept such an account need not be that one ceases to use the thick concept. One can go on making promises and talking of the obligation to fulfill them even after accepting the analysis I offer. Nor is a theory of the kind I am defending reductive in the further sense of singling out one thin concept, such as 'obligation', as the only morally relevant consideration and holding that all moral questions have to be settled by balancing competing claims of this kind. (Williams puts this in terms of the principle, "only an obligation can beat an obligation.")

Williams illustrates this principle with the example of a person who has made a promise to a friend but then finds that he has an opportunity to do something that would be important for some cause in which he believes but which is incompatible with fulfilling the promise. He supposes that the example is set up so that we would agree that the right thing for the person to do would be to break the promise. But, he says, in order for the moral theory of the kind he is criticizing to support this conclusion (or even make it tempting) it has to construe promoting the cause as something the agent has an obligation to do. If it is just something he cares about or thinks important, then it cannot trump an obligation.

I think Williams is right in finding this way of viewing things implausible, but I do not see why it is one that a moral theorist must adopt. In most moral theories I know of (certainly, in my own view, if it counts as a theory) the very understanding of the obligation to keep a promise includes the idea that various factors, not themselves competing obligations, can make it permissible not to do the thing promised. Fulfilling a promise can fail to be obligatory simply because it turns out to have costs to the promisor which were unforeseeable at the time the promise was made and which vastly outweigh any interest that the promisee has in the matter. In order to explain why promissory obligations have this form, we need not assume that would-be promisors have obligations to themselves not to undergo such costs.[14] It is enough to see that this exception is something promisors have reason to want and that promisees could not reasonably object to (indeed, they may even have reason to want it, since it makes people willing to enter into promises that they otherwise would be wary of). More generally, in most moral theories (certainly, the more plausible ones) virtually any obligation or duty will be seen as admitting of exceptions of this kind, ones that are triggered by considerations that need not themselves be other obligations.

---

[14] Or, as Williams suggests in a footnote, that the promisors have an obligation to release them from their obligations—*Ethics and the Limits of Philosophy*, p. 222, note 7

I have so far been describing how an ethical theory might explain the thick ethical concept 'promise', but I think that there is a range of other thick concepts in common use—'murder' and 'lie', for example—that can be accounted for in a similar way. That is to say, there is a range of thick ethical concepts whose force and content is best explained by a theory of right and wrong. Insofar as these explanations are successful, they rebut the charge that ethical theory lacks substantive content.

There are, however, other thick concepts that are not plausibly explained in this way. For example, although some people keep their promises because they believe that it would be wrong (in the sense described by a theory of the kind I have proposed) not to do so, others may keep their word because they believe that breaking it would be dishonorable. Their concept of 'honor' may have little to do with what they owe to others and much more to do with an ideal of character they see as appropriate for a person of a particular class, role, or tradition, and perhaps also with an ideal of personal strength. 'Chastity', which Williams often mentions as an example of a thick concept, may also belong to this class. These concepts are most plausibly understood not as special cases of a general notion of moral right and wrong but rather as deriving (or purporting to derive) their force from other, independent values. Someone who defends a substantive notion of moral right and wrong need not deny that there are such independent values that people can have good reason to be guided by. (This is my answer to the second question mentioned above.)

Even though these freestanding thick concepts do not need to be derivable from the notions of right and wrong that moral theories describe, one important form of critical reflection on such concepts takes the form of asking whether they are consistent with the values that underlie this morality. This is how I would describe the kind of reflection to which Williams refers when he suggests that some thick concepts are properly dismissed as forms of prejudice.

But it is also appropriate to ask, about thick concepts to which there is no such moral objection, whether the ideals they involve are really worth valuing.[15] The result of this further kind of reflective criticism is unlikely to be anything one would call a theory—not even a general, systematic account of what is worth valuing, let alone a set of principles. It does, however, involve stepping back from certain of our evaluative attitudes and looking for "justificatory reasons." This search may involve only a very short step back—we may simply conclude, on reflection, that these ideals are worth valuing. But the step may be longer—we may offer further reasons or even rather complicated arguments in support of them. In either case (as the no-escape thesis entails) the justifications that we come up with will be grounded, ultimately, in substantive judgments of value.

I have argued that the notions of moral right and wrong that moral theory describes have more content than Williams suggests. But in adopting what I have called a substantive strategy for explaining how these notions are reason-giving, and in endorsing the no-escape thesis, I am committing myself to the view that the only possible justification for morality in the narrow sense or for freestanding ethical values rests ultimately on substantive judgments of value. I may therefore seem to be in agreement with Williams's view that deliberation and justification must always start from our evaluative dispositions.

In one sense, the one entailed by the no-escape thesis, I do accept this view. The justification of any ethical judgment must rest on substantive claims about what we have good reason to do, and our acceptance of such a claim reveals a certain "disposition of evaluation" on our part. But Williams's use of the terms 'disposition' and 'local' sometimes seems to suggest another idea, which I would not accept: that the justification of ethical values comes to rest not in certain substantive judgments but in the fact that, for various contingent historical reasons, we *have* these dispositions and are thus disposed to accept these judgments. This distinction is one that Williams himself mentions, as the distinction between "the inside point of view, the view from one's dispositions and the outside point of view of those dispositions," and he goes on to suggest that from the latter, "outside," point of view "there is a sense in which [people's dispositions] are the ultimate supports of ethical value."[16]

It is, of course, trivially true that practices of evaluation depend on people's dispositions—that people who entirely lacked these dispositions would not value anything. But it is equally true, and perhaps also trivially so, that judgments about what makes things valuable can be made only from what Williams calls the inside point of view. (This is just the no-escape thesis again.) And from this point of view (or, rather, from the only such points of view that seem to me to survive critical scrutiny) our dispositions to value things are not, in general, what make them valuable. The fact that one is drawn to a certain value, or that this value has in the past played an important role in shaping one's life, or the lives of those around one, *can* be a reason for continuing to be guided by it. Whether, in a particular case, these facts are good reasons or not is, however, a substantive normative question that can only be answered from "the inside point of view." It is not itself settled by any facts about our dispositions. I hope that Williams would agree.

---

[15] I do not think that this is, at the most fundamental level, a question about whether having a disposition to value these things is part of a "flourishing" life. In at least some cases, a life is a good one because it involves valuing things that are (independently) worth valuing.

[16] *Ethics and the Limits of Philosophy*, p. 51.

T. M. SCANLON

Harvard University

REASONS THIN AND THICK

Recent study in ethics has broadened the scope of the subject, from questions of narrow morality in isolation to questions of what to do and why, and of how to regard people, actions, and other aspects of life. In all this, Bernard Williams has been at the forefront. He turns us from narrowly moral right and wrong to reasons more broadly, and so to right and wrong not in a specifically moral sense, but in the sense of being supported all told by reasons of all kinds. This shift lies at the *thin* end of a spectrum, in the terminology that Williams coined. He has insisted too on *thick* ethical concepts and their centrality, concepts like 'cruel', 'brutal', 'dishonest', 'chaste', 'obscene', and the like, concepts that cultures will not always share. Thick concepts somehow have built into them both a tie to reasons and a tie to empirical classification. Reasons couched in thick terms we might call *thick reasons*.

This essay will be on the structure of ethical concepts; it examines accord and dispute in terms of reasons, thin and thick. What is at issue when we ponder reasons? What is at issue when we disagree over them? Owen's father insists that family tradition gives Owen reason to join the army. When Owen denies this, Williams can demand: What could really be at issue between them? What could Father be asserting and Owen denying? To get at the content of a set of claims, the key is disagreement: to understand concepts as concepts, identify what constitutes agreement or disagreement in terms of those concepts. Williams[1] famously discredits "external" claims of reasons, and his attack succeeds or not, if what I have been saying is right, depending on whether disagreement about "external" reasons is intelligible.

Reasons to act may be couched in plainly descriptive terms, in terms, say, of hurt or money or warmth, or they may be couched in terms that are thick. A strong reason not to torment a cat, for instance, is that it would be *cruel*. The meaning of a term like 'cruel' somehow builds in a tie to reasons. Straight agreement or disagreement in thick terms will not always be in the offing: Oscar Wilde could neither agree

---

[1] "Internal and External Reasons" (1980), reprinted in *Moral Luck: Philosophical Papers, 1973–1980* (New York: Cambridge, 1981), pp. 101–13; and "Internal Reasons and the Obscurity of Blame" (1989), reprinted in *Making Sense of Humanity and Other Philosophical Papers, 1982–1993* (New York Cambridge, 1995), pp 35–45.

0022-362X/03/0006/288–304

nor disagree that a passage was blasphemy; "'Blasphemy' is not one of my words." Here again, disagreement is the key: we interpret a word like 'blasphemy' as having a precise sense only when we settle what it would be to agree or to disagree—or to be in no position to do either, the term not being among one's words. Both with reasons and with thick description, when we ask after the logic of a concept, we are in effect inquiring into agreement and disagreement conditions.

I. REASONS EXTERNALISM

That an act would hurt someone else is reason not to do it, many would claim. If kicking a man would hurt him, that is reason not to kick him—reason in and of itself and not just for what it might bring on the kicker's own head. What could all this mean? We might try interpreting it imperatively: if I tell myself that the pain it would cause the man is reason not to kick him, I am telling myself to weigh his pain against kicking him, to weigh the fact that kicking him would hurt him in favor of desisting. I am telling myself how, in my deliberations, to treat the fact that kicking him would hurt him. Of course, talk here of "deliberations" on whether to kick him is strange, since extreme cases aside, I might want the weighing to be so automatic that the thought of kicking him does not cross my mind. But if you ask me, "Why not kick him?" I do have an answer if I care to give it: "Because it would hurt him." I can endorse my automatic reactions by going through a piece of hypothetical deliberation: asking myself whether to kick him and being overly explicit in my reasoning. I tell myself, "Weigh against kicking him the fact that it would hurt him." This amounts to regarding his pain as reason not to kick him.

This is not yet a full account of what it means to call something a reason. It covers only the first-person present, what I mean when I say that something is reason for *me* right *now* to do a thing or not. The first-person present case, though, can at least serve as a test for any account of what it means to call something a reason.

Bear in mind a familiar and crucial distinction. Adolf Hitler had various reasons for his pact with Joseph Stalin. If we ask, though, "What were his reasons?" we are likely not asking the same question as Hitler did if he asked what reasons he had to make a pact with Stalin. Hitler was asking why to do it; we are asking why he did it. Doing it looked to promote his longterm plan for *Lebensraum* for Germans in the East; that, I take it, was a chief broad reason he had for doing it. But in saying this, I am not saying that this was any reason to make the pact; that would be quite a different claim, and not one I would make. Preventing death and destruction in Russia, on the other hand, might well be reason to make such a pact—but that certainly was not Hitler's reason for making it. Contrast, then, *reason to do* something and one's *reason for doing* it: reasons to do something bear on whether to do it, whereas his reasons for doing it explain why he did it.[2] His reasons for doing it are things he saw as reasons to do it and so acted upon. It is reasons *to do* something that are at issue in debates over *internal* and *external* reasons.

What, then, of the second-person case? I tell you that, in your case, there is reason not to kick me, that the fact that it would hurt me is reason not to do it. In saying this, I am of course advising you, telling you to weigh this fact in favor of desisting.[3] Calling it advice, though, is not by itself informative; we are asking after the nature of advice. How, then, can we extend our first-person account of thinking what reasons there are to the case of offering reasons to you—to the case of advice? One way is this: when I think through reasons for you to kick me or not, we can say I am asking myself what to do if in your shoes. When I voice an answer, I share my thoughts with you; I express a state of mind I have reached. In the previous case, when I announced that your pain was reason for me not to kick you, the state of mind I expressed was this: telling myself, for my actual situation, to weigh this consideration against kicking you. When I now advise that the pain I would feel is reason for you not to kick me, the state of mind I express is the same, but for the hypothetical case of being you, or of being just like you in every way. I am saying, "If I'm about to be you, let me weigh against kicking Gibbard the fact that it would hurt him." Thus we proceed from the first-person case to the second. Thinking what reasons there are for me to do something right now is clearly part of thinking what to do; it consists in thinking what facts to weigh as considerations for and against. Likewise, I am saying, thinking what reasons there are for you to do something is part of thinking what to do as you; it consists in thinking what facts to weigh as considerations pro and con if in your situation.

Thinking of reasons to do things, if this is right, is part of thinking how to live. I can think what to do in contingencies I am not in right now, and that amounts to thinking of what reasons there are to do things in those contingencies. What is it to you, though, my thinking on these matters? Perhaps nothing, if you do not at all trust my judgments about reasons. But likewise with questions of plain fact: my thoughts about where I was yesterday would be nothing to you if you did not at all trust my thinking on where I was. In thinking how to live, though, in thinking about your life and mine, we may do best to put our heads together. You may want help in thinking what to do and how to weigh the considerations. You then want me to think about the same question you are confronting: what to do in your situation and why.

---

[2] Robert Audi makes the distinction in roughly these terms; see "Acting for Reasons," *Philosophical Review*, xcv (1986): 511–46, p. 513.

[3] Williams speaks of a "local advisory model" of reasons—*Making Sense of Humanity*, pp. 208–09.

This account extends to the third person. The Marquis, suppose, has no motivation to keep from hurting me, and no motivation could be aroused by sheer information, vivid awareness, and moves toward consistency. I can still say he has reason not to kick me: kicking me would hurt me, and that, in and of itself, is reason not to kick me. What I then am saying, on the account I have been constructing, is this: suppose I am about to be the Marquis, so that I will have no inkling of intrinsic motivation that Gibbard not suffer. My question now is what to do once so transformed: Kick Gibbard or not? Part of my answer may still be this: to weigh against kicking another person the fact that kicking him would hurt him.

When I move back to the second person and address this thought to the Marquis, the advice I am giving him is sincere: I am saying what I think for the case of being exactly like him. He, to be sure, will ignore me—but it is one question what I am saying and another whether he accepts what I am saying. To accept what I am saying would be to tell himself the very thing I myself am telling him: "In this case, weigh against kicking Gibbard the fact that it would hurt him." Vain hope that he will accept this, but I am being sincere; I myself accept it. We have said, moreover, what it *would* be for him to accept what I am saying.[4]

The account I have been laying out, then, accords sense to reason claims that are "external." Anyone who accepts the account is thus an "external reasons theorist"—or as I will say, a *reasons externalist:* she maintains that there can be a reason to do something even when one's "motivational set" is devoid of anything relevant. Or more precisely, she will think such claims intelligible. The Marquis in the example lacks all motivation to keep from hurting me; he lacks even motivations that are potential in various ways. Still, I assert, he has reason not to kick me, strong reason: that kicking me would hurt me. Whether or not this assertion of mine is true, the account I have offered of reason claims treats this assertion of mine as intelligible. Williams, in contrast, is a *reasons internalist:* it is unintelligible what I could be claiming, he maintains, since I admit that the Marquis lacks all relevant motives.

---

[4] See my *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge: Harvard, 1990), pp. 160–64, for this account of the concept of a reason.

When I say that the Marquis has reason not to kick me, he disagrees. The account I have offered says what is at issue between the Marquis and me: the issue is whether, in his case, to weigh against kicking me the pain that it would cause me. His case, to be sure, includes his lack of all intrinsic motivation to avoid hurting others. Our question, though, is what motives *to have*, and how strongly. This is not a question of psychological fact; it is not the question what motives he in fact has. In his first-person deliberations, the Marquis is asking what to do; he is asking, for his own case, why to kick me and why not to kick me. He is *forming* his motivations. In thinking about reasons and talking with him about them, I myself am addressing the same questions for the same case.

II. CRITIQUES OF EXTERNAL REASONS

In a second article on the subject, "Internal Reasons and the Obscurity of Blame," Williams offers two arguments against externalist accounts of reasons. One hinges on the "interrelation between explanatory and normative reasons." Writes Williams: "If it is true that A has a reason to Φ, then it must be possible that he should Φ for that reason; and if he does act for that reason, then that reason will be the explanation of his acting" (*op. cit.*, p. 39). "What he actually does must be explained by his S," that is, by his "motivational set," as Williams calls it. Suppose, then, my pain is genuinely reason for the Marquis not to kick me. It must then be possible that he should desist from kicking me for that reason—and, as Williams says, if the Marquis "does act for that reason, that reason will be the explanation of his acting."

The Marquis lacks all motivation not to hurt me. Is acting to avoid hurting me, then, not even "possible," in the needed sense? True in fact, alas, he will not shun hurting me, in the case I am imagining—but "will not" does not by itself imply "cannot." Sometimes there is ample reason to do a thing and one still does not do it. Desisting from kicking me, to be sure, would mean changing the Marquis's motivations, but such a thing obtains whenever there is reason to desist from an action but one does not. On the other hand, if the Marquis *did* think my pain strong reason to desist, if he believed what I am saying when I tell him that it is, he would perhaps desist. If he thought there was any reason to desist, perhaps, then he would. In this sense, it is fully possible that he act on this reason.

"If he does act for that reason," Williams adds, "then that reason will be the explanation of his acting." In one sense, something like this does follow from the account I have been offering: that it was the Marquis's reason for desisting would indeed help explain his desisting. My hurt would then be *his* reason for desisting, and this would consist in the fact that he took it to be a reason to desist and that his doing so explained his desisting. Can we burden the external reasons theorist, though, with a stronger reading? Suppose the Marquis, through some amazing transformation, were to stop himself kicking me, and do so on the grounds that kicking me would hurt me. Agreed, the explanation, or part of it, would then be that avoiding hurting others was now among his own reasons for acting. Can we plausibly read the demand, though, as a stronger one, that is, that the explanation cite the finding that not hurting others truly *is* a reason?

In "Internal and External Reasons," his initial article on these subjects, Williams clearly does place the stronger demand on external reason claims. The external reasons theorist, he observes, wants something further: not only "that the agent should acquire the motivation *because* he comes to believe the reason statement," but moreover, that this is "because, in some way, he is considering the matter aright" (*op. cit.*, pp. 108–09). If the agent deliberated rationally, the external reasons theorist must insist, he would come to be motivated. But since ex hypothesi, someone like the Marquis has no motive of decency or compassion to deliberate from, Williams concludes: "I see no reason to suppose that these conditions could possibly be met" (*op. cit.*, p. 109).

One move to meet these conditions, though, is more or less trivial. Rational deliberation, the external reasons theorist can say, is deliberation that "tracks" genuine reasons. It is deliberation that responds to which sorts of reasons obtain in one's situation. Suppose that in the case of the Marquis, my pain indeed is genuine reason not to kick me. Then his deliberation is rational only if he weighs my pain against kicking me. He deliberates rationally only if the reasons to act, in his case, become *his* reasons for acting. As matters stand, to be sure, the Marquis just does not "see it"; he does not see the hurt he knows I would feel as reason not to kick me. If my pain really *is* reason not to kick me, he is blind to this reason. In this regard, however, he is just not deliberating with ideal rationality—so might claim the external reasons theorist.

Can we take seriously this talk of "seeing" reasons? Perhaps it should be laughed aside: it suggests a mysterious Platonism; it suggests that the rational agent responds to the layout of reasons in a nonspatial realm, discerning these objects, being moved by them, and so acting. This is a perceptual model of reasons, on which reasons are something we apprehend and whose apprehension motivates the apprehender. Why should we think there exists any such a nonspatial realm into which we can peer? Neither Williams nor I welcome bald metaphysical claims along these lines. They seem fantastic—and if they *were* true, how could we know it?

We need to ask a further question, though: As people who act and who think about what we are doing, can we coherently deny there is any such a thing as "considering matters aright"? Can we renounce all contrast between thinking rightly and thinking wrongly? We think what to do, and we can also think about thinking what to do, accepting some ways of thinking and rejecting others. In doing this, we are thinking how to think and how not to think. We are forming partial contingency plans for how to plan our lives. I tell myself, from time to time, "I am thinking about this the wrong way." To think such a thing is just to reject the way I have been thinking. Planning how to think and how not to think, then, amounts to arriving at views as to what constitutes thinking aright and what does not, what makes for "considering matters aright." Talk of "seeing what to do" is at least as natural as talk of "seeing what the reasons are," and this talk insinuates a distinction between really "seeing" and the illusion of seeing. We make this contrast when we think how to plan our lives.

It is widely thought that sheer logical coherence is indeed a requirement of planning, that if my plans are logically incoherent, I fail to consider matters aright. All of classical normative decision theory rests on this doctrine; I consider matters wrongly, for instance, if I commit a sunk cost fallacy. As a self-conscious planner, though, must I think I can make mistakes that are substantive, and not just a matter of formal coherence? Can I err by treating as reasons considerations that are really no reasons at all? A snob I rightly despise, imagine, would think badly of me if I liked that song, and I find I care. Can I not think this a mistake? To think it a mistake is to settle on not caring—insofar as my thinking what to care about can control my carings.

There is, I admit, far more to explore to vindicate such an account of the concept of seeing matters aright. Why are questions of how to consider our actions disputable, if they are? I do not at all claim to have gotten to the root of these issues, nor to have answered all worries that Williams should have with the line I am taking. I am suggesting, though, that thoughts of "considering matters aright" are a part of self-conscious planning. I can say, "To consider matters that way is to consider them aright," and that roughly amounts to the imperative, "Consider matters that way!" If we cannot renounce such imperatives, we cannot deny all sense to talk of "considering matters aright."

In his later article "Internal Reasons and the Obscurity of Blame," Williams mounts a further "challenge to the externalist theorist":

"Suppose we take the externalist view, and so accept that it can be true that *A* has a reason to Φ without there being any shadow or trace of that presently in his *S*. What is it the agent comes to believe when he comes to believe he has a reason to do Φ?" (*op. cit.* p. 39). I have nothing direct and informative to say on this; the man just comes to believe that there is reason to do it. No naturalistic synonym for this exists, if the account I have sketched is right. In that regard I agree with nonnaturalists, with philosophers who are nonnaturalists on reasons to act. It would not help, I agree with Williams, to say that the man comes to believe some special nonnatural fact. I do, though, say what this state of coming to believe consists in: for the Marquis to come to believe that my pain is reason not to kick me would be for him to accept a directive to weigh my pain in favor of desisting.

Imagine, says Williams, that a man is brutally nasty to his wife, and nothing we can tell him moves him in the least not to be. Many things can be said, "that he is ungrateful, inconsiderate, hard, sexist, nasty, selfish, brutal," and "that it would be better if he were nicer to her." What, then, does it add to say that he has *reason* to be nicer to her? I have not yet talked about thick terms like the ones that Williams here invokes, and the answer to this challenge will depend on how such terms work. One possibility is this: perhaps if the way the man treats her is "nasty," that entails analytically that he has reason not to treat her the way he does. Thus, if you did not think he had any reason to treat her differently and you were not inconsistent, then you would be misstating your views if you said that his way of treating her was "nasty." If this is how the term works, then of course saying he has reason not to act as he does adds nothing. Once I say that his way of treating her is nasty, I have already committed myself to the less specific view that he has reason not to treat her that way. Note that, if that is so, we shall have to understand talk of reasons to act if we are to understand talk of being "nasty." Reasons to act are then the first thing to explain, even though these other things we can say identify particular kinds of reasons.

Alternatively, these thick terms, or some of them, might not entail what the man himself has reason to do. Instead, they might somehow combine empirical classification with thinking how to feel about the man, how we others have reason to feel about the way he treats his wife. Suppose the term 'nasty' works this way. Then calling him "nasty" will leave another question unaddressed: What considerations, if any, to weigh, if in his shoes, toward treating his wife more considerately? That is part of the question of how to live, applied to the situation of being him.

III. DISAGREEING ON REASONS

If externalist accounts of the concept of a reason must somehow fail, the objections of Williams which I have surveyed do not fully show what goes awry—so I have argued. Now, Williams may take it as obvious that the kind of account I want to offer of what reason claims mean is defective. He assumes that reason facts will be straight, naturalistic facts descriptively couched, and I am offering no such facts. I offer instead the deliverances of thought experiments about what considerations to weigh in someone's case, and how.

Suppose, though, that the thought experiments of two people come out opposed. You consider, hypothetically, being the Marquis who thrills to kicking people, and who does not weigh their pain as in itself counting against kicking them. You say, "If I'm in his shoes and exactly like him, let me weigh the pain of others against kicking them." The Marquis says, nonhypothetically, "Let me not weigh their pain in the least as counting in itself against kicking them." Does this really constitute a disagreement? Why not just say that you and he have different dispositions, two different contingency plans for living? I say that you and he disagree on what considerations to count if in his shoes—but why say this? It seems futile for you and him to argue: true, if he accepted what you say, he would come to count their pain toward desisting. Nothing he already accepts, though, commits him to this. Why dispute such a thing? Why dispute what to do and why in his circumstances? Can there be genuine disagreement on what to do if one is the Marquis?

I raise this question not to dispatch it, but to point out an area that needs exploration. Consider the matter briefly, though: in planning my own life, I cannot well avoid viewing myself as sticking to plans or changing my mind, agreeing or disagreeing with my earlier self; otherwise, my thought lacks all continuity. Now, I am not good at thinking all by myself how to live and why. I want us to put our heads together, letting you try out ways of doing my thinking for me. I let your thoughts enter my stream of thinking as candidate thoughts of my own, to be accepted or rejected. If we take this stance to each other, we are treating the deliverances of each other's thought experiments on how to live as subject to agreement or disagreement.

With the Marquis, however, treating his thoughts as my own would only corrupt me. No pragmatic grounds in this case support thinking together about reasons, treating his thoughts as answers to questions that I ask myself. Now, of course, argumentative impasse threatens even with the plainest of facts. The special creationist and I can reach an impasse on whether the seeming remains in the La Brea Tar Pits stem from real animals and people. That we might reach an impasse, then, in no way shows, all by itself, that no real dispute exists. Still, I face a puzzle: I am proposing to treat planning questions—questions of what to weigh how in one's decisions—as subject to agreement and disagreement among us. For this I offer pragmatic grounds, but these pragmatic grounds can be lacking in some cases. Here is a cluster of issues well worth investigating, and how external reasons would fare in the exploration I must leave unsettled. Instead, I now shift to thick concepts and their tie to reasons.

IV. THICK CONCEPTS AND CLASSIFICATION

A gang of boys torture a cat. Why not join in? Because it is distasteful, no doubt, but above all because it is cruel. And indeed it *is* cruel; in David Wiggins's phrase, "There is nothing else to think." Says Williams: "True, if you use the concept 'cruel', there is nothing else to think but that their behavior was cruel. But there is nothing in the situation, or in the discourse of these other people, that can recruit somebody into using the concept 'cruel' if they don't already."[5] What is at stake in recruiting them into the concept is not just empirical classification. You are recruited only if you oppose what you find cruel, if you think it abhorrent. It is important to ask how this might work. Great powers are claimed for thick concepts, and we should do our best to get to the root of which powers they really have and which they lack. Such terms are bound to be loose and fluid, of course, and sharpening them can only be artificial. If, though, we steer away from explicit analysis, we may be guided anyway by implicit views of how these concepts work, and these implicit views risk being faulty, authorizing runabout inference tickets and the like. We need to ask, then, how such terms could work in our ethical thinking if their meanings were clear and sharp. Williams has some of the most explicit discussions in the literature on how such concepts might work, and there is ample reason to join him in the inquiry.

The concept 'cruel', on Williams's view, is one the boys might understand without making their own. Sheer understanding of our view is not enough by itself to make them regard what they are doing as cruel and abhor it; they can understand our view and reject it. What, then, is missing with them? Even to understand the concept, we may well think, they must see its point, and that requires imagination—but they can see the point and not adopt it.[6] An account of the concept, if all this is right, must yield two different ways the concept could be theirs or ours: comprehending the concept, and standing ready to use it. To stand ready to use the concept, one must of course comprehend it, but not the reverse. Someone who comprehends the concept but rejects it, though, presumably does understand what it would be to stand ready to use it (as a historian might come to comprehend a past notion of chastity she herself does not share).

---

[5] "Truth in Ethics," *Ratio* (new series), VIII, 3 (December 1995): 227–42, p. 237. He adds some qualifications, but what I have to say will not depend on them.

One thing that thick concepts do, according to Williams, is classify. That goes not just for thick concepts whose points are acceptable, but for thick concepts that Williams understands but rightly rejects standing ready to use. When in the days of Augustine, a Christian called an action unchaste, sometimes that person spoke truly and sometimes falsely. If their concept chaste is not ours, however, we ourselves cannot join in with them and judge an act chaste or unchaste in their sense. Augustine calls his earlier life unchaste, and he ought to know; we ourselves can hear him and say "true." But we cannot then add, "Indeed it was unchaste"—for his term 'unchaste' is not one of our words; we can mention it and understand it, but not use it. But Augustine, Williams thinks, was genuinely classifying, whether or not there was good point to classifying that way.

Now, I myself think this wrong: if we reject Augustine's concepts, we cannot coherently credit him with achieving classification. If he speaks truly, then we outsiders can learn to join him in speaking truly; if we rightly judge it improper for us to join him, then we must deny that he is really classifying.

Sometimes Augustine and Monica disagree about what is chaste and what is not; they disagree, say, on whether a woman's exposing her hair to men at large on the street is unchaste. Now if, when they use the term, they are really classifying, then at least one of them is getting it wrong—or else the act is in a penumbra of vagueness that surrounds the term's clear cases. The two agree, however, that exposing hair is not "sort of chaste" and "sort of unchaste"; it is not penumbral. She thinks it clearly chaste, and he thinks it clearly unchaste. If we say that one of them speaks truth and say which one, we are taking sides on this question. But we, I am imagining, reject the point of their classifying acts as chaste or unchaste. Their dispute, then, is one we are unequipped to judge, or to regard as admitting judgment. We might alternatively say that this question of classification is penumbral, that each of them is sort of right and sort of wrong. If we say this, however, we are again taking sides, this time against both of them—for they agree that this is no penumbral matter. We cannot, then, stay out of their disputes and still think them to be classifying, either precisely or vaguely. They as full users treat the question of classification as a real one; we, in rejecting the point of the term, have precluded ourselves from doing so. In rejecting the point of the classification, we forgo the claim that they are achieving a genuine classification.

---

[6] For the history of this point, see Williams, *Ethics and the Limits of Philosophy* (Cambridge: Harvard, 1985), note 7 to chapter 8, pp. 217–18.

Williams disagrees; he proposes a way for an anthropologist or historian to award marks for truth and falsity. The outsider can simulate; he can play act: retaining the ethnographic stance and refusing to adopt their concepts seriously, he can still simulate being a speaker of their language. Call their language *L*: "In that case," says Williams, "he will be following the rule of truthfulness-in-*L*." And that is a rule of truth. Says Williams: "If a person's utterance is true in *L*, and the person is speaking *L*, then the person's utterance is true."[7] To simulate, thinks Williams, the ethnographer recognizes what is true and what is false.

Simulating, though, need not consist in following a rule of truth, if that means attempting actually to speak truths. It consists, we might put it, in *simulating* following a rule of truth. To simulate belief in witchcraft, you simulate classifying people as witches or not. If you simulate well, you respond to the same genuine features of people as does the believer you simulate: social marginality and the like, perhaps. You respond with a simulated classification that matches the earnest beliefs of the subject you simulate. You simulate thinking true that certain people are witches, but the beliefs you simulate, you know well, are false. Likewise, if you want to simulate a person who believes things that are neither true nor false, you aim to simulate regarding those things as true; you do not aim for simulated belief in real truths. The ethnographer who can simulate his subjects, then, need not be following a rule of simulated saying of things he regards as true in their language. He might proceed that way, but simulation does not in general require proceeding that way. He must, of course, respond to those real features to which they respond in their classifying, or the simulation will not be accurate. He need not, though, be responding to the putative features his subjects think they are tracking. Accurate simulation, then, need not, even for the most part, deliver a test for the truth of the beliefs simulated.

---

[7] "Truth in Ethics," p. 239.

In short, to believe a thing to be true is to agree with it, even if you do not join in with the way it is said. With full users of a thick concept, agreement and disagreement depend not just on the facts as a discerning nonuser would see them; they depend also on attitudes or something of the sort. Those who lack the attitudes that would make them users, I am saying, cannot join users in regarding those thick claims as true or false. They do, of course, see users as reacting to refined and genuine features of their social milieu—but they cannot regard the thick terms they reject as classifying by those genuine features.

V. LOADED TERMS THAT DO CLASSIFY

None of this is to claim that no concepts whatsoever work as Williams describes, combining classification and attitude. Racial epithets may sometimes work this way: where the local population stems from different far parts of the world, classification by ancestry can be factual and descriptive, but, alas, the terms people use for this are often denigrating. Nonracists can recognize things people say as truths objectionably couched. With that I agree.

A more interesting case might be the one that Williams has in mind in his discussion of thick concepts. Thick concepts are developed in a local social milieu, and different arrays of concepts fit different social circumstances. Williams's whole purpose in stressing thick concepts is to portray the value of concepts sensitively attuned to refined social circumstances. An ethnographer, then, can arrive on an exotic scene, get to know it, and credit her subjects with thick concepts that are just the ones to use in their circumstances. These are not the ethnographer's concepts at home nor those of her fellows, but her homegrown concepts would ill equip her ethically to live a good life as one of her subjects.

In this case, I agree, she can regard things they say in their thick terms as true, and occasionally as false, without thinking they could be translated into her home language. How, then, can she distinguish this case from the case of vile or defective thick concepts? The ethnographer's question is an ethical one, but in hypothetical form: her questions are how to live if one is like her subjects, and how to think of matters if one is like them. If she thinks her subjects are doing well in living the kinds of lives they live together, her answer to this may be: think as they do. She will not then *just* be simulating them: she will be simulating them among other things, and so be trying to understand what it is like to be them, but she will in addition be making up her own mind for this wild hypothetical eventuality. She is pondering, hypothetically, how to regard matters if in their shoes; she is not just simulating their regardings.

In that case, she has come to speak their language. They are no longer using terms that she is debarred from using, at least in her thought experiments. To be sure, she is still debarred, I am imagining, from using their terms to direct her own life, at home or among them: the terms do not apply at home, and they do not apply to her position as outside observer and ethnographer. They do, though, apply to the hypothetical situation of being one of them, acting and reacting to others. The ethnographer can then agree and disagree with them because she speaks their language, applying it to the same restricted range of situations that she regards her subjects as competent in describing. She ascribes truth and falsehood to them by coming to speak their language to herself.

She will then ascribe knowledge to them, knowledge couched in thick terms. These are terms she has made her own for restricted purposes, and so she can say what it is that they know. Will she think that reflection could destroy this knowledge? If she has reflected and rightly thinks it knowledge, she herself has reflectively come to this knowledge; reflection has not destroyed it.

In *Ethics and the Limits of Philosophy*, Williams contrasted two ways of looking at thick judgments: first, the *objectivist* stance, regarding her subjects as trying to get truth about values, or second, the *ethnographic* stance, viewing their judgments as part of their way of living—in his words, as "a cultural artifact they have come to inhabit" (*op. cit.*, p. 147). Artifacts just viewed as artifacts, though, are not true or false; the ethnographer will not progress to ascribing truth or falsehood until she agrees or disagrees with their judgments, in the same sense in which they agree or disagree with each other. Williams asks whether they could, "in exercising their concepts, express knowledge about the world to which they apply them" (*op. cit.*, p. 147). On the nonobjectivist view, he says, "various members of the society will have knowledge, when they deploy their concepts carefully, use the appropriate criteria, and so on" (*op. cit.*, p. 148). To do this, I agree, they must be responding to some features of the situation, and so perhaps they know about these features. We cannot explain what is going on without attributing some knowledge to them—perhaps that is right. What I deny is something further: that they express this knowledge in their thick claims. Sometimes, they may well do so, I agree, but I deny that we and the ethnographer can judge that they do while taking a purely descriptive, ethnographic stance. When we judge their claims true or false or neither, we are passing to questions of how to live, or how to respond, in their circumstances, to the features they respond to.

Often, the decently cautious course on such questions is agnosticism. We must beware of the arrogance of shoving quick, dogmatic pronouncements down the throats of people who know their own lives intimately. Remember also, however, that we too are human beings, entitled to think about questions of human living—and also that when power is in a person's hands, that brings with it responsibilities to make careful judgments, assess one's knowledge and ignorance, and act accordingly or sit on one's hands. Anthropologists in this century reacted against arrogant Eurocentric practices, but we need not agree entirely with the counterorthodoxy they erected. The ethnographer, I am saying, off duty, can view her subjects as they themselves could were they sufficiently analytic.

VI. THICK DISPUTES

If how people classify with a thick term depends on their attitudes and their attitudes are faulty for their situation, then they do not get it right or wrong with their classifications. It is not just that faulty attitudes attach to the classification; faulty attitudes may fail to produce classifications that are genuine. So I have been claiming: I reject the view that people with faulty thick concepts succeed in marking out a descriptive classification, and then just botch the attitudes to go with that classification.

How, then, do thick concepts work? I will not attempt an answer, but I do want to comment on the job of finding one. First, as I have said, any account of a thick concept is likely to be overprecise. When conceptual troubles rear their heads, we mostly do not sharpen our concepts; we change the subject to something else in the neighborhood. Our question, then, should not be precisely how thick terms work, but whether they could be sharpened to fit their job description, the job description I have been chiefly drawing from Williams. And again, agreement and disagreement are the key to this: a concept is sharply delineated when we know what it is to agree or disagree with applications of it.

Ruth the ethnographer observes this: a story is told in public, and some of her subjects claim that the story was "*kukak*," which she roughly translates as "obscene." Off duty as a purely descriptive ethnographer, Ruth is not agnostic; she has settled how to regard the story in the hypothetical case of being one of her subjects. She does not, imagine, treat the term as vague, or perhaps she is experimenting with a sharp interpretation within an admissible range, as she sees it. Then she can take one of three stances. She can agree, or she can disagree, thinking it true or false that the story is *kukak*. (Notice, I have her adding a word to the English in which she thinks and converses with other observers from outside; in my experience, that is very much the way outsiders talk when learning a culture.) She can also demur; she can do neither, thinking *kukak* judgments defective in a way other than being straight false. In her thought experiments, she may conclude that '*kukak*' is not one of her words for non-make-believe thinking about the lives of her subjects.

Our account of thick judgments, then, should allow for all three of these possibilities. I will not offer an account, but allowing for these possibilities will be part of the job description of such an account.[8]

What, then, ties thick concepts conceptually to reasons? An externalist on reasons can think that any thick concept has reason claims somehow built into it—that, say, by the very meaning of the term, an act is 'cruel' only if it is abhorrent, in the sense that everyone has reason to abhor it and to shy away from committing such acts. Williams must reject any such claim; he thinks that tormenting a cat is cruel even if the boys who torment it lack all reason to shun cruel acts. Perhaps he should say that an act is cruel only if *we* all have reason to shy away in horror from committing such acts—whether or not the cruel boys also do. "We" here might consist of everyone who stands ready to use the concept cruel. Along lines like this, Williams the internalist on reasons may too claim that the concept of a reason is built into any thick concept, and he could say that this is what makes thick concepts distinctive.

I myself am puzzled, though, about how to fill out such an internalist account of thick concepts. Users of the concept 'cruel' debate such things as whether putting a murderer to death without physical pain is cruel. This seems just the sort of issue for which a thick concept like cruel is tailored. What, though, does this disagreement consist in? What is at issue? An externalist on reasons can say this: the issue is whether to avoid and oppose such executions, viewing them with horror. Can the internalist offer an answer? Jack has in his motivational set *S* a horror of all such executions, whereas Jill lacks any such horror for certain cases. Does Jack, then, have reason to oppose such executions and Jill not? They both correctly adhere to internalism on reasons, suppose. Must they each recognize which reasons the other has or lacks? If so, they now agree on who has reason to do what, and they agree on the plain facts of the case: the physical painlessness and psychic torment. I now find it unclear what is left for them to dispute, what further issue an internalist can find conceptually legitimate.

---

[8] I attempt such an account for the term 'lewd' in "Thick Concepts and Warrant for Feelings," *The Aristotelian Society Supplementary Volume* LXVI (1992): 267–83

I end inconclusively, then, in two respects. First, I have attempted no externalist gloss of a concept like 'cruel', and I do not know whether a successful one could be devised. Second, Williams may, for all I have said, have an internalist answer to the puzzle I just raised, even though I myself do not see how it would go. The concepts that Williams calls "thick" are clearly central to many broadly ethical disputes. In actual cases, the points at issue are bound to be vague, but it would good be to know whether clear issues could be identified—whether, say, those who dispute whether painless execution is cruel could find any clear interpretation of the term on which one thinks it cruel and the other not, so that the two are in direct disagreement. The standard of sharpness to aim for is this: that it be made clear what would constitute agreeing that it is cruel, what would constitute rejecting this view, and what would constitute being disqualified from making 'cruel' one of your words.

ALLAN GIBBARD

University of Michigan

ETHICS AND ONTOLOGY

What are we committed to ontologically by our ethical views and commitments? One common temptation of modern philosophy has been to answer: nothing; moral commitments and factual beliefs are in different spheres. It is one of the great achievements of modernity to have sorted this out, as against earlier views of an order of things, or great "chain of being," in virtue of which certain ends held as valid for human beings—fulfilling our natures, or conforming to our place in the "chain," or whatever. Recently, Jürgen Habermas,[1] in espousing his Weber-derived view about the differentiation of rational spheres, has taken a line of this kind.

Habermas's view has the advantage of allowing for our sense that one's ethical views can be right or wrong, better or worse. He espouses a "cognitivist" position, even though what makes the right position right is not its corresponding to "facts." Many of the attempts to establish a fact/value dichotomy have foundered on their espousal of an implausible subjectivism. Emotivism saw value statements as expressions of our emotional reactions to certain objects. "Murder is wrong" had roughly the same content as, "Ugh murder," where the interjection 'ugh' expressed horror, disapproval, rejection.

But this notoriously came to grief as soon as one recognized how much ethical judgments functioned logically like descriptive statements. "Murder is wrong," for instance, can figure in valid syllogisms; for example, "Murder is wrong," "This act is a murder," therefore, "This is wrong." More fundamentally, various ethical predicates: 'courageous', 'generous', 'noble', plainly apply on the basis of fine discriminations, which different speakers can master more or less well. The basic idea underlying emotivism, that the ethical judgment reflects our reaction to a possible or actual state of affairs, which itself could be described in purely "factual" terms, seems just wrong.

There are descriptions of things which pick them out in virtue of their triggering a certain reaction, for example, 'nauseous', where these put no restriction on the ranges of things that can rightly be said to have this property. Something is nauseous if it makes me want to vomit, or if it has this effect on most people, or generally. If someone was made violently ill by some smell, there would be no point in saying: "But this isn't really nauseous."

---

[1] *Theory of Communicative Action*, Volume 1, Thomas McCarthy, trans. (Boston: Beacon, 1984)

0022-362X/03/0006/305–20

But where someone reacts with moral disapproval, or indignation, we often feel justified in criticizing their reaction. This act was not really wrong; you are not justified in reacting to it with indignation. This act does not merit the reaction you have to it.

Bernard Williams[2] has been a key figure in these discussions, in which he has helped lay to rest the early unsophisticated forms of the fact/value dichotomy. But his ultimate position escapes easy definition. I know I disagree with it, but I am not sure I can say just where: perhaps the very definition of a fruitful philosophical difference.

Now as Williams, and also John McDowell,[3] have pointed out, the proper application of moral terms cannot be explained by a separable neutral descriptive component, a set of nonevaluative criteria, which are combined in the ordinary usages of these terms with an evaluative component which could be peeled off, as R. M. Hare[4] has tried to argue. With terms like 'courageous', 'generous', and the like, the concept has the extension that it has in part in virtue of its evaluative force.

What has all this got to do with ontology? Well, a widespread motivation for these attempts to split fact and value has been the need to do justice to (what has been seen as) a crucial feature of the universe. In our post-Galilean age, this has been seen as itself devoid of meaning and value; this in contradistinction to earlier conceptions of cosmic order, such as that of Plato, where the order of things is explained by the realization of a plan which itself has intrinsic value, articulated by the idea of the good, or the principle of plenitude, or some such notion. Value comes into the world with us, the human agents who evaluate. Hence the easy step to theories like emotivism, or views that see value as residing in our reactions to what is in itself neutral reality.

McDowell, whose argument I want to follow here for a while, has been concerned to refute a widespread form of ethical "naturalism," in one meaning of this term. This is the view that arises among thinkers for whom seeing humans as part of nature means seeing their behavior and life form as ultimately explicable in terms that are consonant with modern natural science. The late J. L. Mackie is one protagonist of this position whom McDowell has frequently engaged, as is also Simon Blackburn.[5]

---

[2] See, among other writings, *Ethics and the Limits of Philosophy* (Cambridge: Harvard, 1981).

[3] I shall be drawing mainly on the essays collected in his *Mind, Value, and Reality* (Cambridge: Harvard, 1998); hereafter referred to as *MVR*.

[4] *The Language of Morals* (New York: Oxford, 1952).

[5] Mackie, *Ethics: Inventing Right and Wrong* (New York: Penguin, 1977); Blackburn, "Errors and the Phenomenology of Value," in Ted Honderich, ed., *Morality and Objectivity: A Tribute to J. L. Mackie* (Boston: Routledge, 1985), pp. 1–22.

Now, from this standpoint, there is something problematic about "values," particularly ethical ones. It is hard to see values as part of the ultimate furniture of the universe once one has adopted a post-Galilean stance and no longer understands the cosmos in terms of meaning. So we can be tempted to see our attribution of value to things as a kind of error, systematic and understandable, but error nevertheless. This was Mackie's position. Or we can understand these attributions as a kind of projection we make on the world, as Blackburn does.

McDowell sees this kind of position as deeply confused. There is a sense in which values are not part of the "objective" universe, but this cannot justify our considering them less real for all that. Values are not "in the universe" in the sense that, if we were absent, value attributions could get no purchase on it. That is because an account of what we are attributing, of various value properties, like 'attractive' and 'dangerous', involves at some level or other some reference to their impact on us. (I speak of different levels here, because the impact in question may be on "us" as living beings, or as ethical agents, or as beings who are moved by beauty, or whatever.)

The classical case of properties that are "subjective" in this sense are the "secondary properties" identified in the classical seventeenth-century discussion, like color, heat, taste, and the like. We can say of them, certainly, that they are "not adequately conceivable except...in terms of dispositions to give rise to certain subjective states" (*MVR* 136). An "object's being red is understood as something that obtains in virtue of the object's being such as (in certain circumstances) to look, precisely, red" (*MVR* 133).

But it would be quite unwarranted to speak of error here. The (secondary) quality 'red' would not be there if we sighted beings were not part of the universe; but granted that we are, there are right or wrong attributions of it.

Nor does the notion of a projection seem in place here. Suppose I am angry with you, because I think you have insulted me. You genuinely plead with me that you had no intention to be offensive; quite the contrary, you meant to show respect. But you quite fail to mollify me. At which point a bystander (or maybe you sotto voce) may be judging: "That old codger can't see reason, because he's so full of anger and resentment that he projects hostile behavior onto people even when they are friendly."

Here is a quite different, and stronger, notion of a "subjective" impression. Our hypothesis here is that there was no insult at all, and no hostile demeanor on your part. I "projected" all this erroneously on you, no doubt because of the largely unadmitted anger and hostility boiling inside me.

McDowell's idea is that a major source of this kind of naturalism is a confusion between these two kinds of "subjectivity," moving from the claim that certain properties are inescapably subject-related to the conclusion that they are therefore erroneous, or less than fully real.

When we move beyond secondary properties to values, the issue is further complicated because questions of *merit* now arise (*MVR* 143). Something is really red because it looks red in standard circumstances; there is no further argument. But when I call your act "dishonest," I can be forced to defend myself through chains of reasoning. Maybe you did not know what you were doing; maybe there was a recent price change which I had not heard of, and so on, and so on. Attributions in this domain are sometimes forced to demonstrate their rational grounds.

And it is clear that they can be well grounded. I follow McDowell in taking the example of the dangerous, what is to be properly feared (*MVR* 143 and ff.). Let us start with what is dangerous to life and limb (call this the *L-dangerous*). This is obviously "subjective" in our first sense: no living beings in the universe, no such thing as L-danger. But it is clearly an "objective" matter whether something is really life threatening or not, one in which we can frequently be mistaken, ignorant of key facts, and the like.

This is so clearly the case that we perhaps have to look elsewhere than this confusion to uncover the deep motivations of naturalism. Their problem does not lie with value predicates like 'L-dangerous'. That is because their conception of what it is to be part of nature is being explicable in terms consonant with post-Galilean natural science; and they are confident that biology can be encompassed in such terms. (That is why, indeed, as we shall see below, they sometimes even try to understand ethics in (socio)biological terms.)

But when we come to ethics, it is a different matter. Ethics involves a range of "values" that are essentially understood to be on a different level, to be in some way special, higher, or incommensurable with our other goals and desires. We would not have a category like the ethical or the moral, unless this were so.

Giving an account of this distinction has been one of the most difficult and knotty issues in ethical theory. The famous categorical/hypothetical distinction was Kant's attempt to define it. McDowell gives an interesting critique of this (*MVR*, essay 5). In a sense, Kant fails here because he tries to shirk the task of recognizing certain substantive activities or purposes as higher. He tries to make the distinction turn on formal considerations, on whether or not there is a substantive purpose at play at all. McDowell brings out the qualitative specialness of the ethical in discussing what it is for the requirement of virtue to override those of our other goals and desires. When the courageous person sets aside his desire to run for safety in the face of the enemy, it is not that he sees his goal of defending the polis as *outweighing* that of personal safety. This would be to make two goals commensurable. It is rather that, within the space of questions that define the ethical, this kind of rival consideration is neutralized, counts for nothing.

Remaining within the Aristotle-derived example, we can say that, in the light of the higher goal, *to kallon* or the "noble," saving our skin just does not weigh here. In the deliberation, it is not outweighed by something practically "heavier," but just set aside as irrelevant (*MVR* 47, 92).

(This incidentally opens a new and much more fruitful avenue for addressing the issue of akrasia or "incontinence," as McDowell points out (*MVR* 92–93); the classical puzzle comes from framing the question: How can something rationally assigned a lesser weight end up motivationally heavier? Much of the paradox disappears once we free ourselves from this homogenized image of commensurable weights.)

I could not resist this short digression into McDowell's Aristotle interpretation and ethical theory. But returning to the issue of naturalism, it is clear that this qualitative status of the ethical is a deep source of the trouble. Naturalism, as I have been using it here, which holds to the omnicompetence of post-Galilean science, believes that it can cope with the life-related values: survival, flourishing as a living being, including health, strength, energy, the desires which de facto accompany life: sex, food, (in certain cases) dominance. Let us call these the *L-values*. But it cannot see how values of an incommensurably higher range can have a place in post-Galilean nature.

That certain values should have higher priority over others can be easily understood, for example, survival over some marginal pleasure. That any animal should be "wired" by evolution to respect this priority is easy to accept. Hence desires that are dominant, stronger, more persistent than others must be part of the picture. But the ethical dimension of human life turns on something different: purposes, activities, and qualities which are not necessarily (alas) motivationally dominant, but which are understood and lived as higher in the sense of more worthy. But the higher in this sense is one of the things expunged from the cosmos by post-Galilean science. It had its place in the great "chain of being," but not in the "mechanized" world picture.

Hence, naturalists think, there has to be some error here (Mackie), some projection onto the world (Blackburn), involved in the moral dimension of our lives.

There is a tension between phenomenology and ontology. The former, properly and honestly carried through, seems to show that values of this higher status, like *to kallon* for Aristotle, are ineradicable from our deliberations of how to live. But ontology, defined naturalistically, says that properties of this kind can have no place in an account of things in the world, including its human and animal denizens. These properties are ineradicably "queer," in Mackie's celebrated expression.

One way of resolving the tension is resisting the phenomenology. This kind of position regularly crops up, and one sees it today with certain theories of sociobiology. We try to explain ethics in terms of L-values. Why do we have prohibitions against murder? Why do we value cooperativeness, generosity? It is obvious: the survival conditions of gregarious species like ours require a certain degree of mutual trust and comity. Hume was one of the originators of this kind of account; and the fundamental idea is expressed in the thesis of Philippa Foot, that "moral virtues are qualities necessary if men are to get on well in a world in which they are frightened, tempted by pleasure and liable to hurt rather than help each other" (quoted in *MVR* 177, n21).

Whatever the worth of this as some sort of genetic account, it plainly fails to capture what we mean by virtue. If we try to track our sense of the incommensurably higher through our responses of admiration and its opposite, we can see that whatever the ultimate links with survival and group flourishing, the admirable is never simply defined in terms of them, and sometimes even runs athwart them (*fiat justitia ruat coelum*). We can try to reduce the gap, by paring away at the virtues that we recognize. We repudiate the stern proponent of justice at any price as a "fanatic" or an "enthusiast." But the gap remains: even the most stripped-down utilitarian holds a value like "rationality" as a virtue in the ethical sense; that is, she sees this quality in people not just as useful, not just as desirable, but also as *admirable*. Maybe some day social science will show that having a society of tubby people contributes greatly to mutual tolerance and well being. Tubbiness will then be seen as useful and desirable, but it will not win admiration. That admiration and its opposites are such an ineradicable part of the human life form testifies to the centrality of values that are seen as essentially higher, more worthy.

The gap also shows up if we think of ego's deliberation on what to do. How far could it get on the basis only of considerations of what is in general good for human life? How does the general value of a virtue show that I ought to aim to realize it in my life? McDowell expounds this brilliantly in his parable of the rational wolf (*MVR* 169–73).

Once one sees that phenomenology can not be forced into line with ontology, the other recourse of naturalism is precisely that explored by Mackie and Blackburn. You accept that that is how it looks, but you insist that it is not really like this in the objective world. The difficulty is that this begins to pose epistemological problems for the naturalist ontology. We might set this up in a syllogism in order to pinpoint the difficulty: (1) everything in nature is to be explained in terms of post-Galilean science; (2) humans and their life forms are part of nature; therefore (3) humans, and so forth, are to be explained in post-Galilean terms. (3) then gives the background in terms of which values appear "queer."

But like such syllogisms in Barbara, one can ask what is the evidential relation between major and minor premise. How do we know (1) to be true? Is it because we have independently verified (2), and a whole host of such particulars? Or do we know (1) a priori? This latter claim seems a bit tall, so we would have to fall back on the independent verifiability of (2).

But this is what is now in trouble. One strong bit of evidence for (2) would consist in our being able to make sense of our deliberations in post-Galilean terms. But that is exactly what the phenomenology has shown we cannot do. Failing that, we can appeal to an explanatory account of what we do, including deliberations about higher values, exclusively in post-Galilean terms. But such accounts have nowhere been produced. They are entirely in the realm of promissory notes for a remote future. They are drafts on a bank which may be solvent in 2200, but has not a bean today.

Thus the queerness resides rather in projectivism. It leaves values standing in our actual life-world as inescapable in our real-life deliberations, but then makes a global, marginal comment to the effect that they are not really real. This is a strange claim, not least because it is hard to see where the warrant comes from.

But the strangeness also comes out in the lability of this kind of claim. As it stands, it is purely external to ethical life; that is, it is not a move within ethical reflection and deliberation. But it stands very close to claims that are internal, attempts like the above to pare down the list of acceptable virtues, based on closely related considerations: that we are after all natural, animal beings, that our first allegiance ought to be to life and happiness, and so on. And indeed, the same thinkers very often are found making both kinds of claim: both that values are queer, and that certain values that seem remote and inhospitable to the L-values are just wrong—such as those exalted by ascetic ideals, or a too great concern for one's own integrity. It is not always clear where the argument for one flips over into an argument for the other—or whether the two ultimately conflict. If you play the game internally with such vigor, how seriously can we take your claim that it is all a kind of shadow play?

McDowell's resolution of this tension is to challenge the ontology. Why should we assume that the human world should be explicable exclusively in post-Galilean terms? The existence of human beings brings a whole flock of subjective properties with it. This human world of thick and strong evaluations needs to be understood on its own terms, and not cut and sliced a priori so that it fits the post-Galilean model. This world cannot be made transparent to Galilean science. In relation to this it is as opaque as Stanley Cavell's "whirl of organism" that McDowell loves to quote (see, for instance, *MVR* 206–07). But this can begin to make sense to us if we do not consistently try to denature it.

This fundamental point, about the integrity of the human world, resounds through a number of the dimensions of McDowell's thought; not just on ethics and ontology, as we have seen, but also on issues of explanation, epistemology, and meaning.

From the naturalistic standpoint, it is normal to see the human world as somehow constructed out of Galilean materials. We take in an input that is characterized naturalistically—various forms of energy, or pulses of stimulation, or discharges through neurons. The world of human significance is then constructed in a certain "processing" of this "information." There was originally a certain intellectualist version of this: we construct our view of the world by putting together simple "ideas." But this has now largely been replaced by a reductive-scientistic version that sees the processing as taking place below consciousness, through the functioning of the nervous system and the brain.

Even beyond the disputes in cognitive science, and about Artificial Intelligence, this outlook is deeply anchored in modern thought. It accounts for a certain reading of Ludwig Wittgenstein as a skeptic, which McDowell deconstructs (in *MVR*, essays 11 and 12).

Saul Kripke's[6] famous discussion of what he sees as the basis for the "private language argument" is the best example. Kripke grounds his interpretation on Wittgenstein's remarks about following a rule. Wittgenstein pointed out in the *Philosophical Investigations*[7] that after any number of initial instructions or paradigm examples, it is possible with a little ingenuity to devise a number of incompatible ways of continuing the series or following the rule. Even given a series like: 2,4,6,..., 998, 1000, we could imagine someone going on: 1004, 1008, on the ground that, as the learner understood it, the increments of two were only valid up to a thousand, when they should be upped to four. But we often have quite a strongly-held common view on what is the right way; so that we find Wittgenstein's examples here weird and funny. But how can we know the "right" way?

---

[6] *Wittgenstein on Rules and Private Language: An Elementary Exposition* (Cambridge: Harvard, 1982).

[7] G. E. M. Anscombe, trans. (New York: Macmillan, 1953).

One possibility would be that the person who has grasped the rule has worked out, perhaps largely unconsciously, all the alternative possible interpretations, and has eliminated all but one, which we call the right one. But this clearly will not work, for the number of such possible weird interpretations is indefinite. So then we might flip from the intellectualist to a mechanist stance, and see the person as hard wired so that she follows only one way. The rules are like "rails," as McDowell puts it. But this cannot account for the fact that we live this as something you can get right or wrong; it is following a rule, not an involuntary response to stimulus.

The puzzled conclusion that Kripke arrives at is that nothing in the person's mind, or in her unconscious processing, or in her hard wiring can be what makes it the case that she has learned the rule, and knows how to go on. This "skeptical problem" then is supposed to find in Wittgenstein the "skeptical solution," which attributes the narrowing of endless possibility to the authority of society, demanding conformity from the individual.

It can seem that one is inescapably driven to some such desperate solution as long as we keep thinking of the human agent as an individual information processor in a neutral world. But this is precisely not what a human being in a culture, or life form, or particular "whirl of organism" ever is. For some imagined Martian, or "cosmic outsider" as McDowell has put it elsewhere, it may be an open question whether the increments should be upped to four after 1000, or (why not?) 10,000; or whether we should follow arrows on signposts in the direction of their feathers rather than their tips. Anything is up for grabs. But human thinking is situated thinking, in which any questions that can be raised only make sense against a background or framework of the taken utterly for granted. Our capacity for rational reflection is such that some of what was formerly background can now be put into question, but only against its own background of the unchallenged. To grasp someone's form of life is to understand this pattern of questions against the unnoticed background, perhaps by participating in it unreflectively, perhaps also by some very partial explicit understanding of its limitations.

Where the neutral information processor may stand before endless possibilities, only one may make sense to the encultured human being, whose life form consists of ways of sense making, some of them totally implicit. The enculturated actor does not need an interpretation; she knows right off. That was Wittgenstein's point. The purpose of the weird examples was to offer a reductio of the cosmic outsider picture, not to describe a real problem. This not only makes more sense of Wittgenstein, but it shows once again in another context the advantages of shifting to a view of the human world as to be made sense of on its own terms and not on post-Galilean models.

Does this resolve the issue? Are the ontologically-driven suspicions of "naturalism" toward ethics shown to be unfounded? McDowell seems to have done the trick, and to have reconciled the deliverances of phenomenology (we really discern ethical differences in human life, and these have to be understood as involving incommensurable, higher values), and the basic concerns of a naturalistic ontology, which cannot allow such values into the furniture of the universe. Or put in other terms, he seems to have shown how we can see values as imported into the universe by human beings, without seeing them as merely a matter of subjective reactions or projections which we might peel off things.

I have concentrated on McDowell's work here, but a parallel reconciliation seems to have been achieved by David Wiggins,[8] whose finely calibrated and insightful work has shown how human culture over time can refine our shared perceptions and responses so that we develop socially accepted understandings of the funny, or appalling, or shocking, and the like, which indeed set standards for our individual discriminations.

So has the question been resolved? Are Mackie's error theory and Blackburn's quasirealism just failures to understand the obvious middle ground? Or perhaps Blackburn's present position is just another way of occupying that middle ground?

I am sorry to trouble a hard-won peace, but it seems to me that there remains an issue here. Some sources of the temptation to undercut the phenomenology of the moral life in the name of a post-Galilean "naturalism" have been perhaps laid to rest, but others remain.

One confusion that has been adequately dealt with is that between the two senses of 'subjective'; we no longer need think that because a property is subject-related it must needs be projected onto things, or not properly real, or incapable of yielding a fact of the matter.

---

[8] *Needs, Values, Truth* (New York: Blackwell, 1987), pp 199–200.

And another more subtle and seemingly compelling set of reflections may also have been deprived of its force by the arguments of McDowell and Wiggins, as well as many of the writings of Williams. This is the view I shall call *microreductionism.* Following this, a naturalistic outlook requires that we ultimately be able to explain the "higher" activities of our psychic and social lives in terms of lower order sciences; in the first instance, biological sciences, and then later even in terms of physics and chemistry.

Pursuant to this goal, we can imagine explaining our psychic life in terms of the neurophysiology, endocrinology, and so on, where the explanatory terms and natural kinds of these sciences will in the end prove omnicompetent: they will account for everything, at the "higher" psychic and social levels as well. We imagine the various discriminations we make, including moral ones, being accounted for, say, in a computer-based model of the brain and nervous system, perhaps supplemented with parallel processing. The inputs to the computations that produce such discriminations will be, say, different wave lengths of light impinging on the eye, the stimulation of nerve endings in the fingers, and so on: what W. V. Quine calls "surface irritations."[9]

This is the source of the naturalistic outlook I described a few pages back, which sees the human world as constructed through "information-processing." If we take this path, it can seem that the significance properties of the human world—'dangerous', 'attractive', 'funny', 'appalling', 'noble', and the like—are the stuff of mere "folk psychology," which we have to go beyond to get to the real explanatory factors on the physiological level of input-computation-output. The natural kinds of folk psychology may often be crudely congruent with those of the deeper explanation, but they can also throw us badly off the track, just as our ordinary everyday perception of heavenly bodies can mislead us as to the "sinking" of the sun below the horizon.

A question mark hangs therefore over these significance properties as a class. They can only be declared "real" where they prove congruent with the real underlying factors, and hence they frequently fail of reality. Following this route, moral properties turn out to be dubiously real, not because of any features specific to them, but as one of the broad class of significance properties.

Now, McDowell and Wiggins have done a great deal to combat this microreductionism, and to show how much in our lives becomes unintelligible if we try to sideline the human world. They have not convinced everyone—microreductionism is still alive and well in cognitive science. But let us say we are convinced (anyway, I am), and we either abandon naturalism or adopt a more sophisticated variant. We agree that explanations such as "he convinced them because of his generosity" do not need to be backed on a deeper level with reductive accounts in terms of processing of neutral data.

---

[9] "Epistemology Naturalized," in *Ontological Relativity and Other Essays* (New York: Columbia, 1969), pp. 69–89.

We have now lost a reason to consider all significance properties as "queer" or less than fully "real." But there remain the reaons specific to moral properties, namely, that they invoke incommensurably higher values. These by definition have no place in what I have been calling the "post-Galilean" conception of nature. So even if we no longer aspire to reductive explanations, we still can balk at the full "reality" of the intrinsically higher, in the name of a naturalistic, evolutionary account of the human animal and its life form.

In fact, it is not entirely clear where McDowell and Wiggins stand on the issue of this kind of modified naturalism. It may be significant that the major discussion by Wiggins I referred to above, on the development of shared perceptions and responses of significance properties, centers on the funny as the key example; this is a property which does not at first blush invoke the incommensurably higher as many moral properties do.

So we have not discharged our entire debt to phenomenology once we have laid the ghost of microreductionism to rest. There remains the tension between the phenomenology of the incommensurably higher and a naturalist ontology which has difficulty finding a place for this.

Let us remember the real-life disputes of which these often rather bloodless philosophical debates are echoes. Much of ethics in human history has been articulated in terms of religious or metaphysical beliefs. That is, the incommensurably higher has been understood in terms of, say, God's will for human beings, or the way we have been made by God, or the debt we owe him for being; or else (to take up another major tradition) in terms of the order of things in the cosmos of which we are a part.

Nor were these languages just gratuitous. They made sense to people because they seemed to articulate their sense of the higher. Take one example. One of the core intuition types of our ethical outlook is that which identifies humans as worthy of respect; one might say "incommensurable respect" to mark the notion that their rights are not to be traded away for anything else. This is a particularly strong intuition in our time, but it is part of a family which arguably is always present in human society. Those who command such respect may be a small minority in some societies; they may not include the many subordinate classes—women, slaves, underlings, and so on—but there are always some such. The reasons for such respect vary quite considerably, but it is always there.

Now, the basis of this respect has often been articulated in such terms as the logos that resides in human beings and which is close to the divine in the Universe (some Stoics), or more recently, as our being made in the image of God. Recently, Vaclav Havel[10] has pointed to some such basis for universal human rights.

Is this grounding just a mistake? Or a historically understandable limitation in the imagination of past generations? If so, what kind of mistake was it? Did these people have false views about human psychology? So that they imagined (and others still today wrongly imagine) that you cannot get as worked up about human rights if you do not see humans in this theological or metaphysical spotlight?

But this cannot really be the issue. Because these articulations of the "respect worthiness" of the human did not just offer extra incentives or an extra rhetorical impulse to our inclination to act rightly toward our fellows. They were meant to articulate what respect worthiness consisted in. In other words, they offer further specifications of the concept that is in play here, the one that we can apply rightly or wrongly, and that we therefore cannot see as a simple projection.

And indeed, this emerges clearly in the impact that they have had in the debates about what is due to humans. The theological claim based on Genesis I has been invoked to challenge restrictions on the recognition of rights, as famously with abolitionism in the United States. This is not to say that accepting the theology automatically leads to the emancipatory conclusion; history shows otherwise. But it is undeniable that in this case the abolitionists felt constrained by the logic of their belief. And this led them to propose far-reaching change in the human rights practice of their time. It is hard to fault their logic.

Nor are such issues only to be found in the past. The same theology can inspire action to defend or help people who might be relegated more to the margins in the light of other outlooks. We might think here of Mother Teresa's care for the old and dying in the streets of Calcutta. Or the concern for abortion among many Christians. On the other side, a Utilitarian thinker, like Peter Singer, can countenance infanticide in certain circumstances.

---

[10] "I have often asked myself why human beings have any rights at all. I always come to the conclusion that human rights, human freedoms, and human dignity have their deepest roots somewhere outside the perceptible world. These values are as powerful as they are because, under certain circumstances, people accept them without compulsion and are willing to die for them, and they make sense only in the perspective of the infinite and the eternal. I am deeply convinced that what we do, whether it be in harmony with our conscience, the ambassador of eternity, or in conflict with it, can only finally be assessed in a dimension that lies beyond that world we see around us. If we did not sense this, or subconsciously assume it, there are some things that we could never do"—*The New York Review of Books*, XLVI, 10 (June 10, 1999), p. 6.

Now, there is clearly no simple logic at work here. We cannot take single propositions about the basis of respect and then pair them with universally accepted conclusions about its extension. And it is clear that for any given concrete position on extension, a plurality of ontological views can be arrayed in a coalition behind it. But it is also clear that in the logic of the actual arguments people make and follow, basis articulations have consequences for extension decisions. Some of these arguments are, no doubt, mistaken. But are they all? If not, then the question can arise whether the values we espouse can be supported on the basis of the ontology to which we want to subscribe. Maybe Havel is right, and the kinds of things most of us believe about rights call for an articulation that would take us beyond the bounds of a naturalist ontology. The battle between phenomenology and ontology would break out again.

Or let us take it from another angle. Many of us have a concept of nobility; all the way from Aristotle (translating thus *to kallon*), to E. O. Wilson.[11] Here is a term that invokes the incommensurably higher, which McDowell's argument depends on, according to my reading.

Now one can feel the tension in Wilson's position. If morality is what has been programmed into our brains, if it has "no other demonstrable ultimate function" (*ibid.*, p. 167), then a gap seems to open up between the ontology of what we discriminate in our world (which includes the noble/non-noble), and that of the universe including complex organisms in which we act, and out of which our evolution and present action is to be explained.

One might think that this is just a problem for Wilson, in virtue of the sociobiological account that he espouses; that it disappears once we have taken on board the antireductive arguments of McDowell and others. But it seems to me that a problem still remains.

We give a special importance to nobility. People have been known to risk their lives, their prosperity, their ordinary happiness and fulfillment in its name. It is a powerful ideal. (Of course, it is not for everybody, but for many who espouse it.) How can we explain its rise as a major motivation among this lately developed simian species we belong to? This would seem a legitimate question, although terribly difficult, perhaps unanswerable. But then even though we may no longer acknowledge a necessity to restrict our explanation to the factors that might have impinged on precultural hominids evolving in the late Pleistocene, even though we have rejected a certain kind of evolutionary reductionism, we are still left with a basic conundrum.

---

[11] *On Human Nature* (Cambridge: Harvard, 1978), p. 197. I have discussed Wilson's position in *Sources of the Self* (Cambridge: Harvard, 1989), pp. 406–07.

In terms of the ontology we allow ourselves as post-Galilean naturalists, this strong human motivation, defined intentionally in terms of nobility, has to be explained in terms that recognize no such intrinsic value. That is precisely why the temptation is so strong to explain it in terms of L-values of the kind which figure in sociobiological accounts: for example, there was an evolutionary advantage to warriors feeling they had to risk their lives for the clan, so this very strong impulse got selected for. The response to the current situation is accounted for in terms of our discriminating alternative courses of action as noble/base. But in the genetic account, it is explained in terms of our response to alternatives as strongly/weakly desired. This cannot but evoke the sense that what underlies our strong inclination to nobility is a response that is triggered by a situation that is describable without this value term.

How to avoid this sense of contradiction? Well, by concurring that the evolutionary story does not have to be confined to the extramoral terms and L-values that the post-Galilean restriction requires. Our genetic account has to allow for the emergence of such higher value properties. It has to allow in other words that these properties can figure in their own genetic account, in the sense that they cannot be entirely and adequately explained by the lower-order factors governing the genetic process before their appearance. In terms of L-values, by which we explain the evolution of lower animals, these higher ones represent a kind of surd.

Now, maybe this is what McDowell's antireductive argument is meant to involve. And one could argue that something like this is intended in Wiggins's account of the social-cultural evolution of, for example, the funny. But whatever the views of these writers, it has to be admitted that emergence in this sense puts a severe crimp in the standard naturalism of today's dominant evolutionary theories. We are constantly told that these are meant to explain the rise of the most complex and evolved organisms, including ourselves, by the same principles as govern the emergence of the lowest such beings. Nothing but the chance recombinations of DNA, coupled with the mechanisms of natural selection, is needed to explain the whole saga. But once you accept strong emergence to account for human ethics and culture, then why not also on lower levels as well? One can already hear Daniel Dennett muttering about irresponsible recourse to skyhooks.

In this way, too, it would appear that the hoped-for reconciliation between moral phenomenology and naturalist ontology is, to say the least, somewhat premature.

This much, however, seems clear. We cannot retrieve the original innocent belief of modern, naturalistic philosophical culture that places moral commitments and factual beliefs in different spheres, so that no problem can arise of reconciling phenomenology and ontology. It is clear that naturalism rules out certain bases for ethics, for example, the Stoic or the Christian ones referred to above; and with these, it rules out the property terms that are defined in terms of them. We cannot combine just any set of properties with any old ontology.

The really interesting question that remains is whether the ethic that all of us share, naturalists and antinaturalists alike, say, the affirmation of universal human rights, with their sense that human beings are unconditionally worthy of respect, can consort with the evolutionary naturalism that finds such higher goods "queer." Certainly, the vast majority of our naturalist colleagues believe so. But others, like Havel, seem to have the opposite intuition. Here, on the brink of the really interesting question, I have to break off, partly through lack of time; and partly because the conceptual means at my disposal are still too crude to explore this in an illuminating fashion.[12] I hope to return to this at another time.

CHARLES TAYLOR

McGill University
Northwestern University

---

[12] The conceptual indigence is perhaps not entirely mine. Awareness of these issues is somewhat obscured (a) by the tendency in "analytic" philosophy to neglect questions of articulation in favor of issues around already established formulations, then also more particularly (b) by its tendency to concentrate in ethics on the morality of obligatory action, neglecting issues of the good life and of moral sources; finally, the crunch here is hidden (c) by the understandable confusions as people tack between arguments about the ontological commitments of an already accepted ethic, and revisionist arguments aiming to establish a more stripped-down naturalist ethic (see above, p. 314).

BOOK REVIEWS

Subjective, Intersubjective, Objective. DONALD DAVIDSON. New York: Oxford, 2001. xviii + 237 p. Cloth $45.00, paper $19.95.

Donald Davidson's enormously influential views come to us through articles in the style of the physical sciences. There is no key monograph, no *Word and Object*, no *Naming and Necessity*, no *On the Plurality of Worlds.* In some ways this is a good thing. It certainly helps when we have to convince Deans of Humanities that not every candidate for promotion needs to have a book. Also, it is illuminating to see Davidson approaching one or another crux from different directions in different papers. In any case, this collection and its sister collections—*Essays on Actions and Events,*[1] *Inquiries into Truth and Interpretation,*[2] and the forthcoming volumes, *Problems of Rationality* and *Truth, Language, and History*—are in effect a series of books but (in the case of the published volumes and presumably in the case of those to come) with more overlap than usual between chapters and volumes.

The collection under review gives a comprehensive sense of Davidson's thinking on the topics mentioned in the title. It also contains a final chapter, "Three Varieties of Knowledge," that gives a snapshot of the same (in one of its earlier manifestations, the chapter had the title of the collection), and a very helpful introduction. There are a number of footnotes connecting points made in one chapter to points made in another, or to points made in papers published elsewhere. The earliest paper first appeared in 1982 and the last in 1998.

The themes will be familiar to many readers of this JOURNAL. The denial that there are objects before the mind when we have sensations and think; the analogy between measurement in science (of, for example, temperature) and our use of sentences to give the content of thought; the primacy of truth of sentences as against reference of parts of sentences in the philosophy of language and the consequent (it is argued) relatively harmless indeterminacy of reference; the connection between satisfaction and truth (and an expression of regret that this was originally described as a kind of correspondence theory of truth); a hostility to reductionism combined with a readiness

---

[1] New York: Oxford, 1980.
[2] New York: Oxford, 1984.

0022-362X/03/0006/321–25

to affirm conceptual connections (the latter sits somewhat uneasily with the repeated denial of the analytic-synthetic distinction, or so it seems to me); the thesis that belief and, more generally, thought, requires language; Wittgensteinian views (in the cautious sense of views at least suggested by passages in Wittgenstein) about the connection between our having the concept of an objective world to which we belong and our being part of a community of language users that allows "triangulation"; the intimate connections between interpreting the behavior of others and the contents of their thoughts; the conclusion from the interpretationist approach to content, when given a causal cum externalist gloss, that our beliefs are by-and-large true; and how the fact that one's knowledge of the contents of one's own thoughts does not typically involve a judgment based on one's own behavior sustains first-person authority (not infallibility) about one's thoughts even given externalism about mental content. I found reading the papers end-on-end very illuminating and a reason to buy this volume even for those who already possess the various chapters scattered through other collections or journals.

As well as being enormously influential, some of Davidson's views are very controversial (he adverts to this fact at a number of places in this volume), and, as one reads the volume, there are places where one wishes that Davidson had said a little more by way of addressing or acknowledging that what is being said is very controversial, or about one or another objection that might be (or has been) raised. Davidson is an acknowledged master of expository philosophical prose but there are places where one wishes that he had entered a little more into the hurly-burly of objection and reply. I will give two examples.

In chapter 12, "Epistemology and Truth," he spells out his objections to correspondence theories of truth. A key paragraph runs,

> The correct objection to correspondence theories is .. that such theories fail to provide entities to which truth vehicles (whether we take these to be statements, sentences, or utterances) can be said to correspond. As I once put it, 'Nothing, no *thing*, makes our statements true.' If this is right, and I am convinced it is, we ought also to question the popular assumption that sentences, or their spoken tokens, or sentence-like entities, or configurations in our brains can properly be called 'representations', since there is nothing for them to represent. If we give up facts as entities that make sentences true, we ought to give up representations at the same time, for the legitimacy of each depends on the legitimacy of the other (184).

On the face of it, this is a very radical claim. The argument for saying that sentences, for example, do not represent would appear to apply equally to maps, diagrams of chess positions and so on. But those who look at the chess diagrams in books on Kasparov versus Big Blue or who use street maps to find their way to parties take it for granted that what they have in their hands represents, as it might be, a certain position at a certain stage of the Kasparov match against Big Blue or the layout of streets in Canberra. The same goes for those who create the diagrams and maps. It would have been good if Davidson had gone on to say something like, "You might object that this has the bizarre consequence that, for example, the map of the London underground that visitors to London use to find their way around does not represent the relative locations of the stations. I reply that it does not have this consequence because..."; or something like, "You might object that this has the bizarre consequence that, for example, the familiar map of the London underground that visitors to London use to find their way around does not represent the relative locations of the stations. However, you would be wrong to regard the consequence as bizarre because...." Indeed, as Michael Devitt pointed out to me, some of the questions that need attention are discussed by Stephen Neale in *Facing Facts*.[3] Neale suggests that Davidson might well hold that sentences (and maps and diagrams and beliefs) represent "non-linguistic items" (Neale mentions objects and events as possible candidates) provided they were not thought of as facts or fact-like entities. But in that case, it would seem that sentences and so forth *could* "properly be called 'representations'." Be all this as it may, I think it is a pity that we are left guessing about how Davidson would respond.

My second example is Davidson's famous triangulation argument. This argument is best known through its role in the argument that in order to be a believer, and more generally a thinker, one must be a member of a community of language users; the argument leads to his much discussed conclusion that animals that lack a language (dogs, for example) cannot have beliefs, however natural it is to describe their behavior in terms of belief. The triangulation argument is also part of Davidson's solution to the depth problem, the problem of explaining how it comes to be that our thoughts are about, as it might be, tables and not, say, the impacts of tables on us or the wave patterns of light that go from tables to us. Indeed, in this role, it is a lemma on the way to the conclusion about belief and thought requiring membership in a language community. We will look at the triangulation argument as it appears in chapter 8, "The Second

---

[3] New York: Oxford, 2001, see pp. 72–73

Person," where it introduces Davidson's solution to the depth problem.

Davidson points out there that when we teach the word 'table' to a child by rewarding the child when the word is produced in the evident presence of a table, it is typically the case that the child is being acted upon both by a table and by a surface stimulation pattern. Why not, then, say that the child's "responses are not to tables but to patterns of stimulation at its surfaces, since those patterns of stimulation always produce the response, while tables produce it only under favorable conditions?" (118) In Davidson's view, the answer lies in what we and the child "find it natural to group together" (118) and he goes on as follows:

> We find the child's mouthings of 'table' similar, and the objects in the world we naturally class together that accompany those mouthings is a class of tables. The acoustical and visual patterns that speed at their various rates between ... tables and child's eyes, we cannot easily observe, and if we could we might have a hard time saying what made them similar. (Except by cheating, of course: they are the patterns characteristic... of tables viewed.) Nor do we observe the stimulation of nerve endings of other people.. , and if we did we might find it impossible to describe in a noncircular way what made the patterns relevantly similar from trial to trial....
>
> Involved in our picture there are now not two but three similarity patterns. The child finds tables similar; we find tables similar; and we find the child's responses in the presence of tables similar. It now makes sense for us to call the responses of the child responses to tables. Given these three patterns of response we can assign a location to the stimuli that elicit the child's responses.... It is a form of triangulation: one line goes from the child in the direction of the table, one line goes from us in the direction of the table, and the third line goes between us and the child (118, 119).

The interest of this approach to the depth problem is clear but there is an obvious question to ask about it. In the context of a discussion of the depth problem, there is an appearance of circularity in appealing to what is *found* similar. The question of whether we and the child are finding tables similar, or finding certain stimulation patterns similar, or finding certain acoustical and visual patterns similar is very like the issue under discussion: *finding* Xs to be similar is to have thoughts *about* Xs.

Another way to put the worry is this. There is *a* sense in which the child is responding to stimulation patterns, and to the acoustical and visual patterns Davidson mentions. We state the problem under discussion by noting that there is *another* sense in which the child is responding to tables. That is the sense Davidson is giving to 'respond' in the quotes above; it is the sense in which we say that the child's thought is about tables and not stimulation patterns. But equally, there is *a* sense in which the child finds the stimulation, the acoustical and visual patterns similar, a sense of great interest to certain branches of cognitive science. The solution offered has, therefore, to be understood as giving a sense to 'find similar' on which it is the tables that the child (and we) find to be similar. But what that sense is is very like the problem under discussion.

It would, it seems to me, have been good to find a paragraph starting somewhat as follows: "You might think this proposal is viciously circular; this would be a mistake because...." I am not urging that there is here a big problem for his view. Maybe there is, maybe there is not; it would be absurd to take a position on such a complex matter in a review. But there does seem to be very obviously an issue on which more could have been said, and on which it would have been good to have Davidson's view, or perhaps a footnote to a discussion elsewhere.

FRANK JACKSON

The Australian National University

NEW BOOKS: TRANSLATIONS

HOFFE, OTFRIED: *Aristotle.* CHRISTINE SALAZAR, trans. Albany: SUNY, 2003. xvi, 235 p. Cloth $57.50, paper $18.95; *Categorical Principles of Law.* MARK MIGOTTI, trans. University Park: Penn State UP, 2001. xxx, 311 p. Cloth $75.00, paper $26.50.

HORKHEIMER, MAX, and THEODOR W. ADORNO: *Dialectic of Enlightenment: Philosophical Fragments.* GUNZELIN SCHMID NOERR, ed., EDMUND JEPHCOTT, trans. Stanford: University Press, 2002. xix, 284 p. Paper $21 95.

IRIGARAY, LUCE: *To Speak Is Never Neutral.* GAIL SCHWAB, trans. New York: Routledge, 2002. 277 p. Paper $22.95.

JOXE, ALAIN: *Empire of Disorder.* AMES HODGES, trans., SYLVÈRE LOTRINGER, ed New York: Semiotexte, 2002. 221 p. Paper $9.95.

KANT, IMMANUEL: *Critique of Practical Reason.* WERNER S. PLUHAR, trans. Indianapolis. Hackett, 2002. liv, 284 p. Cloth $37.95, paper $12.95, *Groundwork for the Metaphysics of Morals.* ALLEN W. WOOD, ed. and trans. New Haven: Yale, 2002. xviii, 194 p. Cloth $30.00, paper $12.95; *Theoretical Philosophy after 1781.* HENRY ALLISON and PETER HEATH, eds. and trans., GARY HATFIELD and MICHAEL FRIEDMAN, trans. New York: Cambridge, 2002. xiii, 530 p. Cloth $95.00.

KRISTEVA, JULIA: *Intimate Revolt: The Powers and the Limits of Psychoanalysis.* JEANINE HERMAN, trans. New York: Columbia, 2002. 291 p. Cloth $32.50.

LEFEBVRE, HENRI: *Critique of Everyday Life, Volume II: Foundations for a Sociology of the Everyday.* JOHN MOORE, trans. New York: Verso, 2002. xxix, 380 p. Cloth $35.00.

LEIBNIZ, G.W.: *The Labyrinth of the Continuum: Writings on the Continuum Problem, 1672-1686.* RICHARD T.W. ARTHUR, ed. and trans. New Haven: Yale, 2001. lxxxviii, 484 p. Cloth $75 00.

LORAUX, NICOLE: *The Divided City. On Memory and Forgetting in Ancient Athens.* CORINNE PACHE and JEFF FORT, trans. New York: Zone, 2002. 358 p. Cloth $30.00.

LUHMANN, NIKLAS: *Theories of Distinction: Redescribing the Descriptions of Modernity.* JOSEPH O'NEIL, ELLIOTT SCHREIBER, KERSTIN BEHNKE, and WILLIAM WHOBREY, trans., WILLIAM RASCH, ed. Stanford: University Press, 2002. x, 226 p. Cloth $49.50, paper $21.95.

MARION, JEAN-LUC: *Being Given: Toward a Phenomenology of Givenness.* JEFFREY L. KOSKY, trans. Stanford: University Press, 2002. xi, 385 p. Cloth $70.00, paper $24.95; *In Excess: Studies of Saturated Phenomena.* ROBYN HORNER and VINCENT BERRAUD, trans. New York: Fordham, 2002. xxiii, 169 p. Cloth $40 00; *Prolegomena to Charity.* STEPHEN LEWIS, trans. New York: Fordham, 2002. xii, 178 p. Cloth $40.00, paper $20.00.

MONTINARI, MAZZINO: *Reading Nietzsche.* GREG WHITLOCK, trans. Urbana: Illinois UP, 2003. xxi, 178 p. Cloth $34.95.

PASNAU, ROBERT, ed.: *The Cambridge Translations of Medieval Philosophical Texts, Volume 3: Mind and Knowledge.* New York: Cambridge, 2002. 374 p. Cloth $95.00, paper $35.00.

PATOČKA, JAN: *Plato and Europe.* PETR LOM, trans. Stanford University Press, 2002. xxi, 226 p. Cloth $49.50, paper $19.95.

PENDRICK, GERARD J., ed and trans.: *Antiphon the Sophist: The Fragments.* New York: Cambridge, 2002. xi, 472 p. Cloth $75.00.

PLATO: *The Symposium of Plato: The Shelley Translation.* PERCY BYSSHE SHELLEY, trans., DAVID K. O'CONNOR, ed. South Bend, IN: St Augustine's, 2002. xliv, 95 p. Cloth $20.00, paper $9.00.

ROUSSEAU, JEAN-JACQUES: *Letter to Beaumont, Letters Written from the Mountain, and Related Writings.* CHRISTOPHER KELLY and EVE GRACE, eds., CHRISTOPHER KELLY and JUDITH R. BUSH, trans. Hanover, NH: New England, 2001. xxviii, 334 p. Cloth $65.00.

SACCHI, MARIO ENRIQUE: *The Apocalypse of Being: The Esoteric Gnosis of Martin Heidegger.* GABRIEL XAVIER MARTINEZ, trans. South Bend, IN: St Augustine's, 2002. xix, 146 p. Cloth $28.00.

SAVATER, FERNANDO: *The Questions of Life: An Invitation to Philosophy.* CAROLINA OSPINA ARROWSMITH, trans. Malden, MA: Blackwell, 2002. xiii, 242 p. Paper $21.95.

SCHLEIERMACHER, FRIEDRICH: *Lectures on Philosophical Ethics.* LOUISE ADEY HUISH, trans., ROBERT B. LOUDEN, ed. New York: Cambridge, 2002. xl, 256 p. Cloth $60.00, paper $21.00.

SIMON, YVES, R.: *A Critique of Moral Knowledge.* RALPH MCINERNY, trans. New York: Fordham, 2002. xix, 98 p. Cloth $40.00, paper $20.00.

STEPANIANTS, MARIETTA: *Introduction to Eastern Thought.* JAMES BEHUNIAK, ed., ROMMELA KOHANOVSKAYA, trans. Lanham, MD: Rowman and Littlefield, 2002. xiv, 293 p. Cloth $69.00, paper $24.95.

TODOROV, TZVETAN: *Imperfect Garden: The Legacy of Humanism.* CAROL COSMAN, trans. Princeton: University Press, 2002. viii, 254 p. Cloth $29.95.

VATTIMO, GIANNI: *After Christianity.* LUCA D'ISANTO, trans. New York: Columbia, 2002. 156 p. Cloth $24.50.

VOEGELIN, ERIC: *The Collected Works of Eric Voegelin, Volume 6: Anamnesis. On the Theory of History and Politics.* M.J. HANAK, trans., DAVID WALSH, ed. Columbia: Missouri UP, 2002. i, 438 p. Cloth $39.95; *The Collected Works of Eric Voegelin, Volume 7: Published Essays 1922-1928.* M.J. HANAK, trans., THOMAS W. HEILKE and JOHN VON HEYKING, eds. Columbia: Missouri UP, 2003. 362 p. Cloth $39.95.

VICO, GIAMBATTISTA: *The First New Science.* LEON POMPA, ed. and trans. New York: Cambridge, 2002. lxiv, 302 p. Cloth $65.00, paper $23.00.

VYSHESLAVTSEV, B.P.: *The Eternal in Russian Philosophy.* PENELOPE V. BURT, trans. Grand Rapids, MI: Eerdmans, 2002. xviii, 202 p. Paper $26.00.

NEW BOOKS: PAPERBACK REPRINTS

BALAKRISHNAN, GOPAL: *The Enemy*. New York: Verso, 2000/2002. vi, 312 p. Paper $22.00.

BARRY, BRIAN: *Culture and Equality*. Cambridge: Harvard, 2001/2002. xi, 399 p. Paper $18.95.

BARZUN, JACQUES: *A Stroll with William James*. Chicago: University Press, 1983/2002. viii, 344 p. Paper $18.00.

BERLIN, ISAIAH: *Freedom and Its Betrayal: Six Enemies of Human Liberty*. HENRY HARDY, ed. Princeton: University Press, 2002/2003. xvi, 182 p. Paper $17.95.

BLOOR, DAVID: *Wittgenstein, Rules, and Institutions*. New York: Routledge, 1997/2002. xvi, 173 p. Paper $27.95.

CAVELL, STANLEY: *Disowning Knowledge in Seven Plays of Shakespeare*, Updated Edition. New York: Cambridge, 1987/2003. xvi, 254 p. Paper $23.00.

CHANGEUX, JEAN-PIERRE, and PAUL RICOEUR: *What Makes Us Think? A Neuroscientist and a Philosopher Argue about Ethics, Human Nature, and the Brain*. M. B. DEBEVOISE, trans. Princeton: University Press, 2000/2002. x, 335 p. Paper $17.95.

CULLER, JONATHAN: *The Pursuit of Signs: Semiotics, Literature, Deconstruction*. Ithaca, NY: Cornell, 1981/2002. xxii, 242 p. Paper $17.95.

DE BOLLA, PETER: *Art Matters*. Cambridge: Harvard, 2001/2003. ix, 157 p. Paper $15.95.

EBERSOLE, FRANK B.: *Language and Perception: Essays in the Philosophy of Language*. Xlibris, 1979/2002. xi, 334 p. Paper $23.00; *Meaning and Saying: Essays in the Philosophy of Language*. Xlibris, 1979/2001. xi, 310 p. Paper $23.00; *Things We Know: Fifteen Essays on Problems of Knowledge*. Xlibris, 1967/2001. xi, 381 p. Paper $23.00.

ELDEN, STUART: *Mapping the Present: Heidegger, Foucault, and the Project of a Spatial History*. New York: Continuum, 2001/2002. xiv, 217 p. Paper $29.95.

FÖRSTER, ECKART: *Kant's Final Synthesis: An Essay on the Opus Postumum*. Cambridge: Harvard, 2000/2002. xx, 207 p. Paper $19.95.

GAITA, RAIMOND: *A Common Humanity: Thinking about Love and Truth and Justice*. New York: Routledge, 1998/2002. xxxvii, 293 p. Paper $14.95.

GINZBERG, ELI: *Adam Smith and the Founding of Market Economics*. New Brunswick, NJ: Transaction, 1934/2002. xvi, 265 p. Paper $29.95.

HAMPSHIRE, STUART: *Justice Is Conflict*. Princeton: University Press, 2000/2001. xiii, 98 p. Paper $11.95.

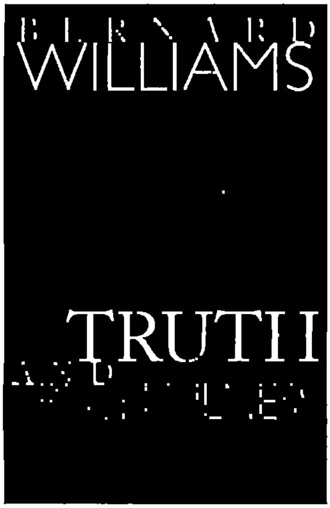
HARTMANN, NICOLAI: *Moral Values*. New Brunswick, NJ: Transaction, 1932/2003. lxiii, 490 p. Paper $39.95.

HEIDEGGER, MARTIN: *Identity and Difference*. JOAN STAMBAUGH, trans. Chicago: University Press, 1969/2002. 147 p. Paper $14.00; *On Time and Being*. JOAN STAMBAUGH, trans. Chicago: University Press, 1972/2002. xi, 84 p. Paper $12.00.

HONDERICH, TED: *Philosopher: A Kind of Life*. New York: Routledge, 2000/2002. x, 441 p. Paper $14.95.

JASPERS, KARL: *Way to Wisdom: An Introduction to Philosophy, Second Edition*. RALPH MANHEIM, trans. New Haven: Yale UP, 1954/2003. i, 208 p. Paper $14.95.

New from Princeton



Truth and Truthfulness
An Essay in Genealogy
Bernard Williams

What does it mean to be truthful? What role does truth play in our lives? What do we lose if we reject truthfulness? No philosopher is better suited to answer these questions than Bernard Williams. Writing with his characteristic combination of passion and elegant simplicity, he explores the value of truth and finds it to be both less and more than we might imagine.

Truth and Truthfulness presents a powerful challenge to the fashionable belief that truth has no value, but equally to the traditional faith that its value guarantees itself. Bernard Williams shows us that when we lose a sense of the value of truth, we lose a lot both politically and personally, and may well lose everything.

"In this exceptionally brilliant book ... Bernard Williams daringly asks—and still more daringly answers—one of the central questions of philosophy: what is the point of telling the truth? Lucid, penetrating, and profound, Williams' reflections are vitally important not for philosophers alone but for anyone interested in human thought and creativity."
—Stephen Greenblatt,
Harvard University
Cloth $27.95 ISBN 0-691-10276-7

OUTSTANDING SCHOLARSHIP FROM CAMBRIDGE

Kierkegaard
A Biography
Alastair Hannay
0-521-53181-0 Paperback $27.00

The Cambridge Companion to the Stoics
Edited by Brad Inwood
Cambridge Companions to Philosophy
0-521-77005-X Hardback $70.00
0-521-77985-5 Paperback $26.00

Essays on Kant's Anthropology
Edited by Brian Jacobs and Patrick Kain
0-521-79038-7 Hardback $60.00

Hellenistic and Early Modern Philosophy
Edited by John Miller and Brad Inwood
0-521-82385-4 Hardback $60.00

Bioethics
Ellen Frankel Paul, Jeffrey Paul, and Fred D. Miller, Jr.
0-521-52526-8 Paperback $25.00

Upheavals of Thought
The Intelligence of Emotions
Martha C. Nussbaum
0-521-53182-9 Paperback $27.00

Under God?
Religious Faith and Liberal Democracy
Michael J. Perry
0-521-82539-3 Hardback $60.00
0-521-53217-5 Paperback $22.00

Facts, Values and Norms
Essays Toward a Morality of Consequence
Peter Railton
Cambridge Studies in Philosophy
0-521-41697-3 Hardback $80.00
0-521-42693-6 Paperback $29.00

Emotions
An Essay in Aid of Moral Psychology
Robert Roberts
0-521-81978-4 Hardback $70.00
0-521-52584-5 Paperback $25.00

Ambiguity and Logic
Frederic Schick
0-521-82458-3 Hardback $55.00
0-521-53171-3 Paperback $19.00

Logic and Theism
Arguments For and Against Beliefs in God
Jordan Howard Sobel
0-521-82607-1 Hardback $65.00*

**Prices subject to change*

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

New from
Oxford

The Roots of Reason
Philosophical Essays on Rationality, Evolution, and Probability
DAVID PAPINEAU
David Papineau presents a controversial view of human reason, portraying it as a normal part of the natural world, and drawing on the empirical sciences to illuminate its workings. In these six interconnected essays he discusses both theoretical and practical rationality, and shows how evolutionary theory, decision theory, and quantum mechanics offer fresh approaches to some long-standing problems.
2003 $29.95

Public Vision, Private Lives
Rousseau, Religion, and 21st-Century Democracy
MARK S. CLADIS
"This is the most important discussion of the conflicts between public and private life yet written by a specialist in modern religious thought. The writing is clear and vigorous, the thinking is careful and well-informed throughout. A wonderful book." —Jeffrey Stout, Professor of Religion, Princeton University
2003 $49.95

Moral Dilemmas
and other topics in moral philosophy
PHILIPPA FOOT
In this new collection, Foot develops further her critique of the dominant ethical theories of the last 50 years, and discusses such topics as the nature of moral judgement, practical rationality, and the conflict of virtue with desire and self-interest.
2003 paper $19.95 cloth $55.00

In Mendel's Mirror
Philosophical Reflections on Biology
PHILIP KITCHER
The articles collected here cover a broad range of topics with similar philosophical and social significance: sociobiology, evolutionary psychology, species, race, altruism, genetic determinism, and the rebirth of creationism in Intelligent Design.
2003 paper $35.00 cloth $75.00

Recreative Minds
GREGORY CURRIE AND IAN RAVENSCROFT
Recreative Minds develops a philosophical theory of imagination that draws upon the latest work in psychology. This theory illuminates the use of imagination in coming to terms with art, its role in enabling us to live as social beings, and the psychological consequences of disordered imagination.
2003 paper $21.95 cloth $65.00

The Oxford Handbook of Aesthetics
Edited by JERROLD LEVINSON
This volume brings the authority, liveliness, and multi-disciplinary scope of the Handbook series to a fascinating theme in philosophy and the arts. Jerrold Levinson has collected 48 brand-new essays, making this the most comprehensive guide available to the theory, application, history, and future of the field.
2003 $110.00

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup-usa.org

OXFORD

World Philosophy

NEW FROM CHICAGO

THE WISDOM OF THE WORLD

The Human Experience of the Universe in Western Thought

Rémi Brague

Translated by Teresa Lavender Fagan

"Rémi Brague is a wondrous scholar in philosophy and philology—two subjects rarely found so impressively together in the same mind. In *Wisdom of the World* his main goal is to show how the world became a guide for human life, and something to be imitated. Beautifully written, his book tells us of the discovery of the cosmos, and gives us some idea of what we have since lost, or deliberately forsaken."—Harvey Mansfield, Harvard University

CLOTH $35.00

ARCHAEOLOGIES OF VISION

Foucault and Nietzsche on Seeing and Saying

Gary Shapiro

"An impressive, challenging, and truly important piece of work. Philosophers aren't, as a rule, used to working so closely with images. Art historians usually aren't so conversant with philosophical theory. *Archaeologies of Vision* bridges the gap by forging a new, hybrid category. Whether they focus on Holbein or Warhol, Shapiro's analyses are sustained, insightful, and a pleasure to read."—Mark Cheetham, University of Toronto

PAPER $27.50

THE ELIXIR OF THE GNOSTICS

A parallel English-Arabic text

Mullā Ṣadrā

Translated by William Chittick

Sadr al-Din Muhammad Shirazi, more commonly called Mullā Ṣadrā (1572–1640), was one of the grand scholars of later-period Islamic philosophy, and has grown to become one of the best known Muslim philosophers.

The underlying theme of Sadra's amplification is emblematic of Muslim philosophy: the importance of self-knowledge in an individual's journey of "Origin and Return," the soul's origins with God and its eventual return to Him. Philosophy, Sadra argues, is the most direct means to self-knowledge—and thus the best tool for navigating this journey.

DISTRIBUTED FOR BRIGHAM YOUNG UNIVERSITY
2 VOLUMES, CLOTH $34.95

THE REFORMATION OF MORALS

A parallel English-Arabic text

Yaḥyā ibn 'Adī

Translated by Sidney H. Griffith

With this text, the tenth-century Syrian Orthodox scholar Yaḥyā ibn 'Adī offered encouragement to the effort promoting moral perfection, especially among kings and other members of the social elite. It now stands as an important Christian contribution, in Arabic, to a strand of moral philosophy that is an integral component of the intellectual tradition of the world of Islam.

DISTRIBUTED FOR BRIGHAM YOUNG UNIVERSITY
CLOTH $34.95

THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, Illinois 60637 • www.press.uchicago.edu

Philosophy for Our Times

Aging, Death, and Human Longevity

A Philosophical Inquiry

CHRISTINE OVERALL

"For those who believe that philosophy should concern real issues of everyday life that genuinely matter to our ability to live well, this book is essential."

—Joan Tronto, author of *Moral Boundaries*

$44.95 hardcover

Plato's *Parmenides*

Translated and with Commentary by

SAMUEL SCOLNICOV

"An immensely clear and interesting introduction and way into the *Parmenides*."

—Georgia Warnke, author of *Justice and Interpretation*

A JOAN PALEVSKY BOOK IN CLASSICAL LITERATURE

$49.95 hardcover

NEW IN PAPERBACK

Henry David Thoreau and the Moral Agency of Knowing

ALFRED I. TAUBER

"A revealing and comprehensive study of Thoreau."

—Eric Wilson, author of *Emerson's Sublime Science*

$24.95 paperback

Caring

A Feminine Approach to Ethics and Moral Education

NEL NODDINGS

SECOND EDITION, WITH A NEW PREFACE

"Noddings examines . . . how caring ultimately functions in an educational context."

—*New York Times*

$19.95 paperback

Graceful Simplicity

The Philosophy and Politics of the Alternative American Dream

JEROME M. SEGAL

"A marvelously textured analysis of the elusive ideal of simple living."

—David Shi, author of *The Simple Life*

$16.95 paperback

The Splendid Feast of Reason

S. JONATHAN SINGER

"Inviting and lively."

—*Publishers Weekly*

$16.95 paperback

At bookstores or order 1 (800) 822-6657 • www.ucpress.edu

UNIVERSITY OF CALIFORNIA PRESS

Important Questions. Thoughtful Answers.

Philosophy is often criticized for being irrelevant to social issues of wide public concern. Not *Philosophy & Public Affairs*! In every issue it demonstrates the relevance of contemporary philosophical thought. It is a journal in which philosophers with different viewpoints, along with philosophically inclined writers from many academic disciplines, bring their distinctive methods to bear on problems that concern all readers—scholars and citizens alike.

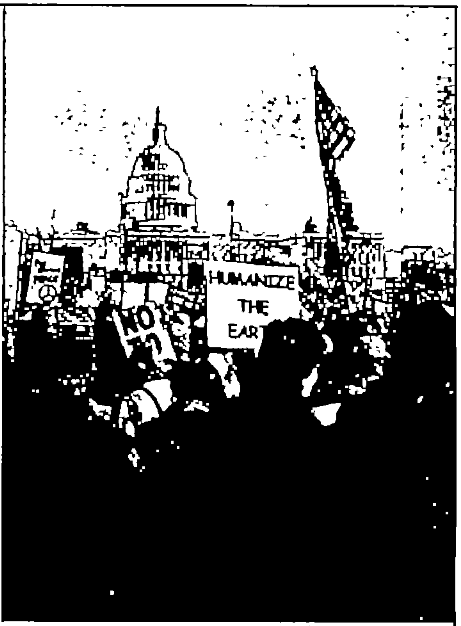


Photo by Declan McCullagh

THE BEST IN CONTEMPORARY THOUGHT

Philosophy & Public Affairs presents important work by today's most thoughtful writers.

Recent articles include.

Samuel Freeman, "Illiberal Libertarians. Why Libertarianism Is Not a Liberal View"
Barbara Herman, "The Scope of Moral Requirement"
Samuel Scheffler, "What Is Egalitarianism?"
Edna Ullmann-Margalit and **Cass R. Sunstein**, "Inequality and Indignation"
Bernard Williams, "From Freedom to Liberty: The Construction of a Political Value"

Don't miss out. Subscribe today.

Philosophy & Public Affairs is published quarterly by Princeton University Press
and distributed by The Johns Hopkins University Press
2003 Annual Subscription Rates (U.S.) Individuals, $40.00; Institutions, $80.00;
Students, $20.00. Canada and Mexico, add $5.00. Outside North America, add $8.50

How to Order: PHONE: 1-800-548-1784 (Outside the USA and Canada call 410-516-6987)
FAX: 410-516-6968. EMAIL: jlorder@jhupress.jhu.edu
www.press.jhu.edu/press/journals/pap
A Project Muse participant

Philosophy & Public Affairs: The Best in Contemporary Thought

A CRITIQUE OF MORAL KNOWLEDGE

Yves R. Simon

Translated by **Ralph McInerny**

This long-awaited book is the first English-language edition of Simon's first book, *Critique de la connaissance morale* (1934) Not only does this work clarify the first stages of Simon's intellectual career, it is also a major contribution to moral philosophy.

90 pp. • 0-8232-2103-2 / $40.00 cloth • 0-8232-2104-0 / $20.00 paper

BEING AND SOME TWENTIETH-CENTURY THOMISTS

John F. X. Knasas

Knasas calls on the tradition established by Gilson, Maritain, and Owen, to build a case for Existential Thomism as a valid metaphysics. This book is a comprehensive discussion of the major issues and controversies in neo-Thomism, including issues of mind, knowledge, the human subject, free will, nature, grace, and the act of being

416 pp. • 0-8232-2248-9 / $65.00 cloth

HEIDEGGER: THROUGH PHENOMENOLOGY TO THOUGHT

Fourth Edition

William J. Richardson

With a new introduction by the author

Preface by **Martin Heidegger**

One of the most frequently cited works on Martin Heidegger that, now in paper, belongs on any short list of Continental philosophy classics Includes as a preface the letter that Heidegger wrote to Richardson and a new writer's preface and epilogue

840pp. • 0-8232-2255-1 / $50.00 paper

Perspectives in Continental Philosophy, 30

RELUCTANT THEOLOGIANS

Franz Kafka, Paul Celan, Edmond Jabès

Beth Hawkins

Beth Hawkins focuses on the problematic faith in the works of Kafka, Celan, and Jabès to reevaluate the notions of God and covenant in light of Nietzsche's "death of God"

240 pp. • 0-8232-2200-4 / $35.00 cloth • 0-8232-2201-2 / $20.00 paper

Studies in Religion and Literature, 4

REINTERPRETING RAHNER

A Critical Study of His Major Themes

Patrick Burke

"Burke's impressive study shows why Rahner remains such an important figure in contemporary Catholic theology This thought provoking book deserves to be widely read" —Paul D Molnar, St John's University

320 pp. • 0-8232-2218-7 / $45.00 cloth • 0-8232-2219-5 / $20.00 paper

At bookstores

FORDHAM UNIVERSITY PRESS

www.fordhampress.com • 800-247-6553

New from **Columbia**

Time Passing
Modernity and Nostalgia
Sylviane Agacinski
Translated by Jody Gladding

"This book traces connections leading from the forbidding landscape of today's political and technological challenges to its deepest foundations in time. Weaves together cultural, philosophical, and literary insights with uncommon verve and lucidity."
—Kevin Newmark, Boston College

240 pages • 12 photos • $29.50 cl
[illegible]

Desolation and Enlightenment
Political Knowledge After Total War, Totalitarianism, and the Holocaust
Ira Katznelson

A major intellectual history of Hannah Arendt, Robert Dahl, Richard Hofstadter, Harold Lasswell, Charles Lindblom, Karl Polanyi, and David Truman, detailing their postwar effort to reclaim the West's moral bearing

"Brilliant, an event not only in the history of ideas, but in our own history, too."
—E. J. Dionne, Jr, author of *Why Americans Hate Politics*

208 pages • $27.50 cl
[illegible]

The Education of John Dewey
A Biography
Jay Martin

"Martin's biography is a good, well-paced read."
—Richard Rorty, *New York Times Book Review*

"Truly terrific ... It is the definitive biography of John Dewey." —Jo Ann Boydston

592 pages • 50 photos • $35.00 cl

NOW IN PAPERBACK
Between East and West
From Singularity to Community
Luce Irigaray
Translated by Stephen Pluhacek

"What happens when a distinguished French feminist philosopher and psychoanalyst takes yoga lessons?"
—*Library Journal*

208 pages • $17.50
[illegible]

EDINBURGH UNIVERSITY PRESS
From Kant to Lévi-Strauss
The Background to Contemporary Critical Theory
Jon Simons

An accessible basic guide to critical post-Enlightenment European thinking, this introductory work treats fifteen key figures in modern Continental philosophy.

272 pages • $24.00 pb

COLUMBIA UNIVERSITY PRESS
columbia.edu/cu/cup 800-944-8648